DOCUMENT RESUME

ED 230 628 TM 830 490

AUTHOR	Huynh, Huynh; Casteel, Jim
TITLE	Technical Works for Basic Skills Assessment Programs. Final Report.
INSTITUTION	South Carolina Univ., Columbia. School of Education.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.; South Carolina State Dept. of Education, Columbia.
PUB DATE	Mar 83
GRANT	NIE-G-80-0129
NOTE	121p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Achievement; *Basic Skills; Educational Diagnosis; *Educational Research; High Schools; Measurement Techniques; *Minimum Competency Testing; Remedial Instruction; Research Methodology; Scores; *Statewide Planning; Student Evaluation; *Testing Programs; Test Interpretation
IDENTIFIERS	*South Carolina Basic Skills Assessment Program

ABSTRACT

This report deals with ways to report basic skills test data which would facilitate the identification of student weaknesses. Under study are the technical aspects and methods associated with the reporting of objective-referenced data. An exploration is then made into the use of patterns of errors in responding to basic skills test items to possibly improve various score reporting processes. In addition, the feasibility of using these patterns to construct instructionally equivalent test forms is discussed. Finally, an approach is presented to project budget requirements and allocation of resources in school districts or states in which instructional remediation is a corollary of a basic skills assessment program. This work is geared to the needs of planners of statewide or districtwide basic skills assessment programs and to other people such as students, parents, and teachers who would benefit from test interpretations which are detailed yet simple. Procedures which enhance the identification of weaknesses in the acquisition of basic skills, particularly among disadvantaged students, will undoubtedly contribute to the mission of testing for instructional purposes and for program evaluation. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

ED230628

TECHNICAL WORKS FOR BASIC SKILLS ASSESSMENT PROGRAMS

Huynh Huynh
Jim Casteel

FINAL REPORT

March, 1983
Educational Research Program
College of Education
University of South Carolina
Columbia, South Carolina 29208

This research was supported by the National Institute of Education, Department of Education, Grant NIE-G-80-0129, and by the Office of Research, South Carolina Department of Education

TM 830 490

CONTENTS

Acknowledgements v

Abstract vii

PART A: FOCUS OF THIS STUDY: THE SOUTH CAROLINA BASIC SKILLS ASSESSMENT PROGRAM

1. A Glance at the South Carolina Basic Skills Assessment Program 3

PART B: REPORTING OBJECTIVE-REFERENCED TEST DATA

2. A Comparison of the Rasch and Two-Parameter Logistic Models in the Context of Decisions Made for Each Objective in a Basic Skills Assessment Program 27

3. A Minimax Approach to Setting Multivariate Passing Scores for Subtests when the Passing Score of the Total Test Is Known 41

4. Reporting Test Scores as Percent of Correct Responses and Construction of Unit Items in a Pool 53

PART C: EXPLORING THE USE OF PATTERNS OF INCORRECT RESPONSES

5. Exploring the Use of Patterns of Incorrect Responses in Score Reporting Via the Bock Multinominal Latent Trait Model 61

6. Exploring the Use of the Log-Linear Model in the Identification of Group Differences in Patterns of Incorrect Responses 77

PART D: CONSIDERATIONS FOR BUDGET ALLOCATION

7. Assessing the Budgetary Impact of Remediation in Basic Skills Assessment Programs: I. Statistical Considerations 87

8. Assessing the Budgetary Impact of Remediation in Basic Skills Assessment Programs: II. Instructional Consideration 101

PART E: SOME VIEWS ON PSYCHOMETRIC ISSUES

9. A View on Future Psychometric Issues in Mental Measurement . 117

ACKNOWLEDGEMENTS

The research reported herein was supported in part by the National Institute of Education. It was also conducted under the auspices of the Office of Research of the South Carolina Department of Education as the technical basis of the Basic Skills Assessment Program (BSAP). During the funding period from September, 1980, through December 31, 1982, Lawrence Rudner, in his capacity as Project Officer, was very helpful in smoothing out the many details which otherwise would have hampered the research activities.

Anthony Nitko spent countless hours reading and commenting on most of the first draft of this final report. Lifa Yu helped with some of the data analysis. Sarah P. Seaman-Huynh and Martha Reynolds provided the extensive analysis of errors in the reading items. The Office of Research of the South Carolina Department of Education, Paul Sandifer, Director, and Vana Meredith, Supervisor, was most cooperative in supplying the BSAP data. To all, our special note of thanks.

Several procedures described in this report were conceived and worked out in conjunction with the development and implementation of the South Carolina BSAP. The complete trust of Paul Sandifer and Vana Meredith and their staff in our work was most appreciated.

Though the frontier of research and knowledge is unlimited, the number of hours available for research during weekdays for a faculty member with a full teaching load is not. Sarah Seaman-Huynh was graceful about her husband's long absences on many weekends.

The final report was typed by Michele Davis Bergen. Her superb typing deserves more than a mere note of appreciation. Finally, gratitude is extended to the Computer Services Division of the University of South Carolina, which made available hardware and software to facilitate the completion of this project.

Huynh Huynh
Jim Casteel

ABSTRACT

Recently several school districts and states have implemented programs testing for minimum competency in the basic skills. The test data are to be used to diagnose a student's deficiency and to provide for instructional remediation. Several technical and practical issues related to these monitoring programs are discussed and solutions are provided in this report.

The first part of this report deals with ways to report basic skills test data which would facilitate the identification of student weaknesses. Under study are the technical aspects associated with the reporting of objective-referenced data. An exploration is then made into the use of patterns of errors in responding to basic skills test items to possibly improve various score reporting processes. In addition, the feasibility of using these patterns to construct instructionally equivalent test forms is discussed. Finally, an approach is presented to project budget requirements and allocation of resources in school districts or states in which instructional remediation is a corollary of a basic skills assessment program.

This work is geared to the needs of planners of statewide or districtwide basic skills assessment programs and to other people such as students, parents, and teachers who would benefit from test interpretations which are detailed yet simple. Procedures which enhance the identification of weaknesses in the acquisition of basic skills, particularly among disadvantaged students, will undoubtedly contribute to the mission of testing for instructional purposes and for program evaluation.

PART A

FOCUS OF THIS STUDY: THE SOUTH CAROLINA BASIC SKILLS ASSESSMENT PROGRAM

CHAPTER 1

A GLANCE AT THE SOUTH CAROLINA BASIC SKILLS ASSESSMENT PROGRAM

1. Introduction

In attempting to reverse the decline in the level of student achievement over the last decade, several states have implemented statewide testing programs assessing minimum competency in the basic skills. Many of these programs aim to insure that high school graduates possess a minimum level of academic achievement and have acquired the skills required to function effectively as adults in American society by requiring high school students to pass an examination. When used in this manner--that is, as high school exit examinations--minimum competency examinations do not have the positive connotation of some other basic skills assessment programs such as the one implemented in the State of South Carolina. This program is specifically designed for continuous monitoring of the acquisition of basic skills (namely, reading, writing, and math) across successive grade levels. The results of this type of continuous monitoring program are used to diagnose a student's deficiencies in the basic skills and to provide for instructional remediation.

The purpose of this introductory chapter is to provide an overall description of the South Carolina Basic Skills Assessment Program (BSAP) and some of its major technical works. It is within the framework of the BSAP that the research supported under the auspices of the National Institute of Education was conducted. The NIE works will be described in detail in the subsequent chapters.

2. A Brief Description of the South Carolina BSAP

On July 14, 1978, the South Carolina Legislature enacted legislation establishing the South Carolina BSAP. The program is aimed at the establishment of statewide educational objectives in the basic skills (namely, reading, writing, and math) along with minimum

standards of student achievement for kindergarten through grade twelve. The program consists of two separate testing components. First, a *readiness test* is to be administered to all public school students at the beginning of grade one to assess the student's readiness to begin the formal school curriculum. The results of the readiness test are to be used to provide appropriate developmental activities in the first grade. In addition, the school district is to advise the parents of any student not indicating readiness for first grade work to secure a complete physical examination of that child. Second, *criterion-referenced tests* are to be developed in reading and math for grades one, two, three, six, and eight and *writing exercises* for grades six and eight. The purpose of these tests is to diagnose student deficiencies and to aid in determining instruction needed by the student in order to achieve the minimum statewide standard established for each grade level. (An adult functional competency test is also to be administered at the end of grade eleven.)

Readiness Testing

For beginning first graders, the readiness test chosen was the Boehm/Slater Cognitive Skills Assessment Battery (CSAB) published by Teachers College Press of Columbia University. The selection was made in conjunction with the identification of the kindergarten objectives. The readiness test was field tested in the spring of 1979 using a sample of kindergarten students. Prior to testing, the kindergarten teachers' judgements on the readiness of the students were also solicited for the purpose of setting the passing score. Since no longitudinal data were yet available in 1979 on the CSAB for South Carolina first graders, judgements by a cross-section of South Carolina kindergarten teachers were used as a proxy for the actual performance of first graders during the school year. The cutoff score was set at 88 out of a maximum of 117.

Basic Skills Assessment

With full participation of all parties concerned with public education in the state, the South Carolina basic skills objectives in reading and math were identified. These objectives were deliberately formulated to be broad in scope, yet still measurable. In addition, they were so selected that, with effective instruction, the objectives could be achieved. Thus sensitivity to instruction was a major factor employed in the framing of each objective.

The objectives in reading for each of the grades one, two, three, six, and eight are stated in six categories: decoding and word meaning (DW), main idea (MI), details (DF), analysis of literature (AL), reference usage (RE), and inference (IN). In math, the objectives are clustered in five categories: operations (OP), concepts (CN), geometry (GE), measurement (ME), and problem solving (PS).

The development of the reading and math tests was contracted with the Instructional Objective Exchange (IOX), Los Angeles, California. Test items were field tested in the spring of 1980, and the first forms were administered statewide in 1981. For each subsequent year, new forms are developed and administered. As planned, all test forms have items of similar content; in addition they share a number of common items. This was deliberately done so that variations in item characteristics and student ability can be observed from year to year.

3. Setting Passing Scores: Descriptions of Three Approaches to a Set of Data

There are a variety of ways to set passing scores for a basic skills assessment program or minimum competency test. Most procedures can be classified either as content-based or data-based. Variations of content-based procedures have been proposed by Nedelsky, Angoff, and Ebel; they typically focus on some type of subjective judgement regarding the content of items or objectives to be measured by the test and expected performance of an examinee at the borderline of achievement.

Data-based procedures for standard setting, on the other hand, use the examinees' item responses. Most of them rely on an external classification of examinees in *contrasting groups* and seek passing scores which are, in some sense, consistent with the external classifications.

In the context of the South Carolina BSAP tests for grades one, two, three, six, and eight, the setting of passing scores was based on a combination of student responses and teacher judgements. Since all standards are judgemental, the credibility and fairness of those who make the judgement determine the extent to which the resulting passing scores are acceptable to the public. For the BSAP tests, it was felt that teachers who had been teaching the students for almost a year would be in the best position to make judgements regarding the performance of students in the academic areas under study.

During the May 1981 statewide BSAP testing, samples of approximately 3000 students were selected for each of grades one, two, three, six, and eight and for each of the areas of reading and math. A few weeks prior to testing, teachers were asked to classify each student's achievement in each subject area as *Adequate* or *Non-adequate*. In the case of uncertainty, the student was to be classified in the category of *Undecided*. Table 1 reports the descriptive data regarding the achievement in reading for the groups *Adequate*, *Non-adequate*, and *Undecided;* the corresponding data for math are compiled in Table 2. For all grades and subject areas, the BSAP means and medians are in the expected direction; that is, for each situation the *Non-adequate* group has the smallest mean or median and the *Adequate* group has the highest mean or median. Thus there is a high degree of relationship between BSAP test scores and teacher judgements. Since the BSAP tests are deemed to have adequate content validity, this level of correlation indicates that teacher judgements were made on a basis similar to the content measured by the test. It may be recalled that these judgements were made independently of the test scores.

TABLE 1

Descriptive Statistics for Teacher-Judgement Samples,
May 1981 Statewide Testing, Reading Tests

Grade	Subject	Statistics	Combined sample*	Non-adequate group	Undecided group	Adequate group
1	Reading	N	2923	892	194	1779
		Mean	26.08	19.13	22.85	29.99
		Median	27	18	22	32
		SD	7.76	5.95	5.97	5.91
2	Reading	N	2675	862	136	1636
		Mean	26.80	19.83	24.39	30.68
		Median	30	18	25.5	33
		SD	7.92	6.82	6.71	5.62
3	Reading	N	2725	1025	96	1537
		Mean	27.57	22.42	24.24	31.24
		Median	30	23	26	33
		SD	7.17	6.99	7.13	4.66
6	Reading	N	2677	1012	117	1422
		Mean	24.36	18.50	23.01	28.54
		Median	25	18	24	30
		SD	7.56	6.27	4.83	5.58
8	Reading	N	2624	824	135	1626
		Mean	24.40	17.99	24.76	27.68
		Median	26	17	26	29
		SD	7.84	6.80	7.05	6.19

*Including students with no recorded teacher judgement.

TABLE 2

Descriptive Statistics for Teacher-Judgement Samples,
May 1981 Statewide Testing, Math Tests

Grade	Subject	Statistics	Combined sample*	Non-adequate group	Undecided group	Adequate group
1	Math	N	2923	589	161	2125
		Mean	25.32	20.69	24.05	26.74
		Median	27	21	25	28
		SD	4.16	4.61	3.90	2.87
2	Math	N	2672	629	139	1866
		Mean	25.87	22.71	25.09	27.00
		Median	27	23	25	28
		SD	3.72	4.25	3.02	2.85
3	Math	N	2722	838	105	1714
		Mean	22.65	19.29	21.55	24.38
		Median	23	19	22	25
		SD	4.70	4.42	4.53	3.87
6	Math	N	2681	1057	124	1437
		Mean	16.61	12.21	16.56	19.97
		Median	16	12	17	20
		SD	6.34	4.68	5.07	5.42
8	Math	N	2631	1040	140	1418
		Mean	13.41	10.11	14.71	15.73
		Median	12	9	14.5	15
		SD	6.48	4.74	6.64	6.55

*Including students with no recorded teacher judgement.

Three approaches were considered in the setting of passing scores via teacher judgements. They are subsequently described as the *Contrasting Group* procedure, the *Equal Percent Failing* procedure, and the *Undecided Group* procedure.

Contrasting Group Procedure

In this procedure the group *Undecided* is ignored, and the passing score is chosen to be the test score at which a maximum *number of students* are correctly classified. Let N_1 $(x < c)$ be the number of *Non-adequate* students with scores less than c; let N_2 $(x \geq c)$ be the number of *Adequate* students with scores of at least c. Then the passing score is the value c at which the sum $N_1(x < c) + N_2(x \geq c)$ is the highest.

Equal Percent Failing Procedure

The *Equal Percent Failing* procedure focuses on the proportion of *Non-adequate* students and seeks a passing score which yields a similar proportion of statewide students who would fail the test. Since all test score distributions are *discrete,* the *Non-adequate* proportion (based on teacher judgements) and the proportion of students who fail the BSAP test usually cannot be made exactly equal. However, since the BSAP aims at helping *Non-adequate* students, it would make sense to err in the direction that would help to identify these students; hence if two consecutive test scores may be used as the passing score, the higher one would be the more appropriate choice.

Undecided Group Procedure

Another feasible way to set passing scores for the BSAP tests is to focus on the *Undecided Group* and to set the passing score as the median score of this group. This practice presumes that the category *Undecided* is comprised of students who are on the borderline between adequacy and non-adequacy; and the typical *Undecided* examinee would be right on the cutoff score separating students who pass the test from those who fail it. The median is preferable to other

summary measures such as the mean because of its resistance to outlying observations which are common in statewide testing programs.

4. Setting Passing Scores: Results from Three Approaches

Tables 3-5 present the passing scores compiled from each of the three procedures previously described. Along with the passing scores, other descriptive information is also provided. This information is listed under Columns 4-7 and is described as follows.

Column (4): Statewide percent of failing students

Column (5): Percent of failing in *Non-adequate* group (one type of correct decision)

Column (6): Percent of passing in *Adequate* group (another type of correct decision)

Column (7): Percent of correct decisions

There is an additional column in Table 4.

Column (8): Percent of *Non-adequate* students based on teacher judgement

It may be noted that the *Equal Percent Failing* procedure results in passing scores which are equal to the corresponding *Contrasting Group* passing scores in one situation and higher in the remaining nine situations. Except for one case, reading in grade three, the *Undecided Group* passing scores are at least as high as the *Contrasting Group* passing scores.

Except for the math test of grade eight, all three procedures appear to provide passing scores which are intuitively defensible. The passing scores of 11 and 12 provided by the *Contrasting Group* and *Equal Percent Failing* procedures for the math test of grade eight appear too low considering that, with four options per item, the mean chance score is 7.5 and the standard deviation is 2.4. The passing score of 15 provided by the *Undecided Group* procedure seems more acceptable.

In the remainder of this introductory chapter as well as in all subsequent chapters, the *Undecided Group* passing scores will be used for various illustration purposes. (They will be referred to as statewide passing scores or standards.)

TABLE 3

Passing Scores Based on the Contrasting Group Procedure and Relevant Descriptive Statistics

Grade (1)	Subject (2)	Passing Score (3)	Percent Statewide Failing (4)	Percent Failing in Non-adequate (5)	Percent Passing in Adequate (6)	Percent Consistent Classifications (7)
1	Reading	22	30	69	89	83
2		24	34	72	89	83
3		28	40	71	84	79
6		23	41	74	85	80
8		20	28	60	89	79
1	Math	22	16	56	93	85
2		23	19	42	92	80
3		20	27	54	88	77
6		15	42	70	82	77
8		11	38	62	74	69

TABLE 4

Passing Scores Based on the Equal Percent Failing Procedure and Relevant Descriptive Statistics

Grade (1)	Subject (2)	Passing Score (3)	Percent Statewide Failing (4)	Percent Failing in Non-adequate (5)	Percent Passing in Adequate (6)	Percent Consistent Classifications (7)	Percent Non-adequate (8)
1	Reading	23	34	73	86	82	33.3
2		25	36	74	87	82	34.5
3		28	40	71	84	79	40.0
6		24	45	78	81	80	41.6
8		22	35	69	80	79	33.6
1	Math	24	26	70	87	83	21.7
2		25	31	63	83	78	25.2
3		22	39	68	79	75	32.8
6		16	47	76	78	77	42.4
8		12	43	69	69	69	42.3

TABLE 5

Passing Scores Based on the Undecided Group Procedure and Relevant Descriptive Statistics

Grade (1)	Subject (2)	Passing Score (3)	Percent Statewide Failing (4)	Percent Failing in Non-adequate (5)	Percent Passing in Adequate (6)	Percent Consistent Classi-fications (7)
1	Reading	22	30	69	89	82
2		26	38	77	84	82
3		26	33	61	89	78
6		24	45	78	81	80
8		26	49	83	69	74
1	Math	25	32	77	82	81
2		25	31	63	83	78
3		22	39	68	79	75
6		17	53	81	73	76
8		15	57	83	54	66

5. Overall Procedure for Item Calibration

At a very early phase of development of the BSAP tests, the Rasch model was chosen as the general framework for all technical works. The decision was made primarily because the Rasch model is the logistic model which is most consistent with the tradition of using the number of correct responses as the test score. For each test administered in 1981, all items were calibrated on samples of approximately 2600 students each. These are the samples used in the setting of passing scores (see Section 3). The mean difficulty of items in each test was (arbitrarily) set at zero; these items defined a common ability scale for all the subtests covering the objectives. (As may be recalled, there are six objectives in reading and five objectives in math.) The results of the Rasch calibration for the reading tests are reported in Table 6 and those for the math tests are listed in Table 7.

For items which were not part of the 1981 test forms, the Rasch difficulty values were obtained from the field test data collected

TABLE 6

Rasch Item Difficulty Values for Reading in 1981

Objective	Item Sequence	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
DW	1	-1.365	-2.313	-1.991	-0.405	0.505
	2	-1.446	-2.953	-1.053	-1.223	0.370
	3	-1.704	-0.394	-0.452	-2.268	-1.271
	4	-0.994	-1.578	0.610	0.527	-1.523
	5	-1.682	0.571	-0.638	-1.511	-0.280
	6	-1.169	0.482	-0.226	0.197	0.168
MI	7	0.547	0.034	0.371	1.476	0.189
	8	0.260	0.316	0.463	0.425	-0.302
	9	-0.300	0.843	0.845	0.617	0.349
	10	-0.324	1.066	0.761	0.353	0.000
	11	-0.561	0.965	0.857	0.579	0.293
	12	-0.450	0.953	0.543	0.506	0.747
DE	13	0.615	-0.072	-0.160	-0.387	-0.181
	14	0.976	-0.098	-0.885	-0.910	0.293
	15	0.824	-0.084	-0.188	-0.447	-0.275
	16	0.598	0.249	0.306	-0.591	-0.092
	17	0.580	-0.250	-1.192	-0.126	-0.750
	18	0.959	-0.777	-0.788	0.091	-0.783
AL	19	1.199	0.726	-0.646	1.124	0.927
	20	1.103	0.492	-0.700	0.822	-0.106
	21	0.461	0.199	-0.130	1.036	0.027
	22	1.169	0.064	-0.692	0.850	0.753
	23	0.776	0.721	2.327	1.330	0.825
	24	1.645	0.726	1.797	1.097	0.948
RE	25	-1.279	-0.716	0.810	-0.531	-0.710
	26	-0.216	-0.323	0.179	-1.204	-0.167
	27	-0.433	0.352	-0.142	-0.964	0.315
	28	-0.665	-0.298	0.082	-1.524	-0.361
	29	-0.292	0.124	0.239	-1.175	-0.891
	30	-0.068	-0.150	-0.801	-0.392	-0.748
IN	31	0.139	-0.164	-0.268	0.267	-0.205
	32	0.376	0.132	0.234	0.538	0.244
	33	0.262	-0.044	0.506	-0.209	0.319
	34	-0.090	0.185	0.130	0.785	0.325
	35	0.226	0.482	-0.571	0.024	0.687
	36	0.320	0.532	0.461	0.323	0.364

TABLE 7

Rasch Item Difficulty Values for Math in 1981

Objective	Item Sequence	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
OP	1	0.918	0.772	-0.115	-1.274	-0.285
	2	0.475	0.659	0.765	0.023	-0.479
	3	0.710	0.582	0.497	-0.721	-0.471
	4	0.972	2.288	-0.016	-0.084	-0.623
	5	1.168	0.713	0.612	-0.304	-0.316
	6	0.938	-0.054	1.179	1.122	-0.176
CN	7	-1.618	1.970	-0.177	1.037	0.782
	8	0.029	-0.449	0.821	0.675	1.621
	9	-1.053	0.135	-0.658	-0.570	-0.135
	10	-1.174	-0.427	0.377	-0.120	-0.144
	11	0.409	0.495	1.064	1.074	-1.065
	12	4.012	-0.019	-0.507	-1.406	0.638
GE	13	-0.106	0.251	0.041	-0.848	0.281
	14	0.540	0.445	-0.138	-0.107	0.290
	15	-2.711	-0.675	-1.265	-0.195	-0.434
	16	-2.007	-1.507	-1.516	0.665	0.699
	17	0.318	-1.799	-1.094	1.045	0.928
	18	0.253	-0.490	0.415	1.926	-0.392
ME	19	0.230	0.797	-1.003	-0.607	0.791
	20	0.645	0.795	-0.766	-0.179	-0.578
	21	-0.284	-0.414	-0.006	-0.011	-0.762
	22	1.241	-1.667	-1.791	0.734	-0.069
	23	-1.547	-2.147	0.501	-0.176	0.657
	24	-1.817	1.696	1.793	0.965	-0.595
PS	25	-0.176	-0.965	-0.473	-0.200	0.603
	26	-0.106	-0.551	-0.264	-0.663	0.512
	27	0.406	-0.408	0.388	-0.870	-0.488
	28	-0.258	0.772	0.175	-0.237	0.237
	29	-0.454	-0.337	0.471	-0.872	-0.891
	30	0.051	-0.403	0.688	0.143	-0.137

in 1980. At this pilot test administration, three test forms were assembled at each grade level for reading (Forms R1, R2, and R3), and another three test forms were put together at each grade level for math (Forms M1, M2, and M3). Form R1 contained all items (including those subsequently used in the 1981 test forms) in the two objectives of main idea (MI) and decoding and word meaning (DW), Form R2 contained all items in details (DE) and analysis of literature (AL), and Form R3 contained all items in inference (IN) and reference usage (RE). As for math, Form M1 consisted of all items in concepts (CN) and operations (OP), Form M2 consisted of all items in geometry (GE) and measurement (ME), and Form M3 consisted of all items in problem solving (PS). The number of students who responded to each pilot test form in each grade ranged from 282 to 439 with an average of 405 in the reading area. As for the math subject, the number of examinees ranged from 262 to 461 with an average of 395. (The field test design also included Forms R4 and M4, which consisted respectively of items taken from each reading objective and from each math objective. However, due to the availability of the statewide 1981 data, student responses to Forms R4 and M4 were not needed in the item calibration process.)

At each grade and for each subject area, Rasch item calibrations were carried out separately for the three pilot test forms. By use of appropriate sets of linking items, the Rasch difficulty values of all items not included in each 1981 test form were then positioned on the ability scale defined by the items which constituted the 1981 test form. The linking items were selected from the set of items which appeared on both the 1981 test form and each of the three pilot test forms. Two criteria were used in the selection of the linking items. First, the linking items must not show gross departure from the Rasch model. Second, in the bivariate plot of the two estimates of Rasch difficulty levels (one based on 1980 field test data and the other based on 1981 statewide test data), the linking items had to stay close to a regression line with unit slope.

6. Conversion from Raw Scores to Scale Scores

When expressed in raw test scores, the statewide passing scores do not remain the same for all tests. In addition, test security necessitates the use of different forms each year. Although every effort is made to insure that these forms are comparable both in content and in difficulty, there is no guarantee that raw test scores from comparable forms are strictly equivalent. Taking these factors into account, it was felt that a common scale score system would be the best way to express the student achievement in various subjects across various grade levels. In a testing program where items are already calibrated, it is possible to set a common scale score system for all test forms. Although it is a matter of arbitrary decision, the 1981 BSAP test scores are reported on a scale score system in which the statewide passing scale score is held at 700 for all situations; in addition, the standard deviation is set at 100.

Latent trait models may be used in the construction of scale scores for any test. Let θ be the latent trait (ability) for an examinee and $P(\theta)$ be the item characteristic (operating) curve for an item. Then $P(\theta)$ is the probability that the said examinee will answer the item correctly. For a test with L items, each with the item characteristic curve (icc) $P_i(\theta)$, $i = 1,2,\ldots,L$, the test characteristic curve (tcc) is the sum

$$E_L(\theta) = \sum_{i=1}^{L} P_i(\theta). \qquad (1)$$

This is the number of correct responses to be expected from an examinee with ability θ.

On an L-item test, the raw score (number of correct responses) is an integer on a scale extending from 0 to L. For a raw score r, let θ_r be the ability on the ability continuum defined by the test. For the raw scores of $1,2,\ldots,L-1$, the ability θ_r is the solution θ_r of the equation $E_L(\theta_r) = r$. Strictly speaking, when $r = 0$, $\theta_r = -\infty$ and when $r = L$, $\theta_r = +\infty$. To avoid having a scale score of infinity,

one may linearly extrapolate the tcc at θ_1 to get the value θ_0 and at θ_{L-1} to get the value θ_L. Linear extrapolation yields the ability

$$\theta_0 = \theta_1 - 1/E_L'(\theta_1) \tag{2}$$

and

$$\theta_L = \theta_{L-1} + 1/E_L'(\theta_{L-1}). \tag{3}$$

In these formulae, E_L' represents the derivative of $E_L'(\theta)$ with respect to θ.

For the special case of the Rasch (one-parameter logistic) model, the icc is given as

$$P(\theta) = e^{\theta-\delta}/(1 + e^{\theta-\delta}) \tag{4}$$

where δ is the difficulty of the item. For this case, we have

$$E_L'(\theta) = \sum_{i=1}^{L} P_i(\theta)\cdot\left(1 - P_i(\theta)\right); \tag{5}$$

hence $1/E_L'(\theta_1)$ is the square of the standard error of measurement at θ_1 and $1/E_L'(\theta_{L-1})$ is the square of the standard error of measurement at θ_{L-1}.

Let $c(\theta)$ be the cutoff ability and $\sigma(\theta)$ the standard deviation of the ability distribution derived from each BSAP test administered in 1981. For each test in each grade level, the scale score for the raw score r (=0,1,2,...,L) is given as

$$\text{scale score} = 700 + 100\left(\theta_r - c(\theta)\right)/\sigma(\theta). \tag{6}$$

In subsequent statewide BSAP test administrations, new test forms will be assembled for each grade and in each of the areas of reading and math. Each form corresponds to a tcc; this curve provides a way to convert each raw score r into an ability θ_r. Once this is done, the formula $700 + 100\left(\theta_r - c(\theta)\right)/\sigma(\theta)$ will be used to determine the scale scores for the new test form. The cutoff ability and standard deviation $c(\theta)$ and $\sigma(\theta)$, computed from data of the 1981 statewide BSAP, will be held constant across all new test forms.

7. Scale Scores Conversion for the 1981 BSAP Tests

The Rasch item difficulty values for the BSAP tests administered in 1981 were previously reported in Tables 6 and 7. The statewide frequency distributions established at the raw score level are reported in Tables 8 and 9. The constants $c(\theta)$ and $\sigma(\theta)$ for each test are listed in Table 10. Tables 11 and 12 present the scale scores for the 1981 BSAP tests. As may be recalled, for each test at each grade level, the scale scores are linear transforms of the Rasch abilities; the constants defining the transformations are set up so that, for 1981, the passing score is 700 and the standard deviation is 100.

8. An Historical Note

The passing scores based on the *Undecided Group* procedure (Table 5) were recommended as statewide passing scores for the South Carolina BSAP in the memorandum dated October 23, 1981, from Huynh Huynh to Dr. Paul Sandifer. Dr. Sandifer was director of the Office of Research of the South Carolina Department of Education. After lengthy discussions within the department, the passing scores were recommended to the South Carolina State Board of Education, which adopted them in the meeting of March 10, 1982. They were finally passed to the South Carolina Legislature, which had 120 days to voice rejection of the recommended passing scores. Without any formal rejection within the 120-day period, the recommended passing scores became legal statewide passing scores. (Due to fluctuation in the difficulty of items used in subsequent years, all the raw passing scores were located at 700 on the scale scores; the passing score of 700 has become the legal statewide passing score for all BSAP tests.)

9. An Early Trend in Student Performance on the BSAP Tests

Table 13 reports the percent of students in grades 1, 2, 3, 6, and 8 who met the statewide passing score of 700 for the school years of 1980-81 and 1981-82.

TABLE 8

Statewide 1981 Raw Score Frequency Distribution
Reading

Score	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
0	2	1	0	7	10
1	5	0	4	2	11
2	5	1	3	12	9
3	9	2	3	20	19
4	10	2	5	34	41
5	15	7	5	72	103
6	19	23	16	141	190
7	48	49	66	229	314
8	81	91	141	356	450
9	192	194	233	557	667
10	334	325	359	640	775
11	464	560	547	828	863
12	622	803	690	914	1032
13	807	1120	819	1144	1037
14	1015	1334	849	1221	1101
15	1183	1447	936	1346	1164
16	1320	1471	971	1501	1279
17	1509	1366	943	1512	1317
18	1582	1322	914	1610	1429
19	1680	1181	934	1684	1421
20	1713	1095	998	1729	1454
21	1708	1017	1023	1813	1545
22	1722	970	1030	1842	1632
23	1628	976	1155	1832	1708
24	1617	1060	1269	1818	1661
25	1493	1107	1378	2015	1839
26	1497	1150	1575	2074	1807
27	1435	1330	1715	2047	2019
28	1521	1574	1988	1988	2061
29	1539	1689	2272	2105	2198
30	1642	2031	2636	2214	2359
31	1876	2362	3020	2297	2394
32	2058	2891	3546	2267	2683
33	2509	3338	3910	2224	2568
34	3514	4030	4190	2050	2516
35	4186	4259	3882	1676	2096
36	4953	3391	3002	862	1201

TABLE 9

Statewide 1981 Raw Score Frequency Distribution
Math

Score	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
0	4	0	1	11	17
1	1	0	0	17	71
2	1	2	1	57	235
3	2	1	1	148	598
4	0	0	3	332	1163
5	3	1	5	716	1733
6	3	5	9	1130	2424
7	6	12	17	1487	2811
8	12	25	58	1799	2875
9	36	29	97	2070	2988
10	54	56	202	2203	2825
11	93	97	329	2358	2499
12	151	120	523	2422	2425
13	234	185	780	2413	2173
14	268	246	1044	2437	2033
15	444	378	1295	2410	1883
16	559	485	1578	2546	1891
17	744	640	1844	2349	1777
18	899	859	2249	2452	1653
19	1155	1002	2455	2416	1575
20	1437	1197	2681	2244	1537
21	1685	1508	2990	2169	1458
22	1986	1910	3135	1918	1408
23	2357	2251	3384	1771	1327
24	2862	2906	3517	1689	1156
25	3536	3446	3677	1454	1094
26	4537	4253	3760	1139	1052
27	5521	5082	3728	971	841
28	6741	6366	3475	766	653
29	7264	6825	2707	514	467
30	4853	5634	1435	213	185

TABLE 10

Cutoff Points $c(\theta)$ and Standard Deviations $\sigma(\theta)$ of Ability of Students in the 1981 BSAP Administration

Subject	Constants	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
Reading	$c(\theta)$	0.542	1.099	1.076	0.834	1.033
	$\sigma(\theta)$	1.728	1.619	1.563	1.382	1.399
Math	$c(\theta)$	1.963	1.924	1.156	0.292	-0.009
	$\sigma(\theta)$	1.507	1.371	1.181	1.176	1.238

TABLE 11

BSAP Scale Scores for 1981
Reading Tests

Raw Score	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
0	381	314	318	275	285
1	442	383	385	352	359
2	485	432	433	407	412
3	512	463	462	441	444
4	532	486	483	466	468
5	548	505	501	487	487
6	562	521	516	504	503
7	574	534	529	520	517
8	585	547	541	534	530
9	595	558	551	547	542
10	605	568	562	559	553
11	614	578	571	571	563
12	622	587	580	582	573
13	631	596	589	592	583
14	639	604	598	602	592
15	647	612	606	612	601
16	654	620	614	622	609
17	662	628	622	632	618
18	670	635	630	641	627
19	677	643	639	651	635
20	685	651	647	660	644
21	692	658	655	670	653
22	700	666	663	680	662
23	708	674	672	690	671
24	716	683	681	700	680
25	724	691	690	711	690
26	733	700	700	722	700
27	743	709	710	733	711
28	752	720	721	746	722
29	763	731	734	759	735
30	775	743	747	774	749
31	788	757	762	791	765
32	804	773	780	810	784
33	824	793	802	835	807
34	850	821	832	867	839
35	893	866	881	921	891
36	953	930	949	996	965

TABLE 12

BSAP Scale Scores for 1981
Math Tests

Raw Score	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
0	230	201	200	278	331
1	303	279	290	367	415
2	356	336	354	430	475
3	389	371	394	469	512
4	415	397	424	498	539
5	436	419	448	522	561
6	454	438	469	543	580
7	470	455	488	561	596
8	485	470	505	577	612
9	499	485	521	593	626
10	512	498	536	607	639
11	524	511	550	621	652
12	536	523	563	635	664
13	547	536	577	648	676
14	559	548	590	661	688
15	570	559	603	674	700
16	581	571	616	687	712
17	592	583	629	700	724
18	603	595	642	713	736
19	614	608	656	727	748
20	626	621	670	741	761
21	639	634	684	756	775
22	652	649	700	772	789
23	666	664	717	789	805
24	682	681	735	808	822
25	700	700	756	829	841
26	721	722	779	853	864
27	749	749	809	883	891
28	786	784	848	923	928
29	848	841	912	988	989
30	933	920	1001	1078	1074

TABLE 13

Percent of Students Meeting Minimum Statewide Standards
in 1981 and 1982

Subject	Year	Grade 1	Grade 2	Grade 3	Grade 6	Grade 8
Reading	1981	70	62	67	55	51
	1982	72	69	69	62	52
Math	1981	68	69	61	47	43
	1982	68	64	68	51	41

PART B

REPORTING OBJECTIVE-REFERENCED TEST DATA

CHAPTER 2

A COMPARISON OF THE RASCH AND TWO-PARAMETER LOGISTIC MODELS IN THE CONTEXT OF DECISIONS MADE FOR EACH OBJECTIVE IN A BASIC SKILLS ASSESSMENT PROGRAM

1. Introduction

As explained in the introductory chapter of this final report, the South Carolina Basic Skills Assessment Program (BSAP) consists, in part, of reading and math tests to be administered to public school students near the end of grades 1, 2, 3, 6, and 8. Each reading test focuses on six objectives: decoding and word meaning (DW), main idea (MI), details (DE), analysis of literature (AL), reference usage (RE), and inference (IN). Each math test measures student performance in five objectives: operations (OP), concepts (CN), geometry (GE), measurement (ME), and problem solving (PS). For each test there are six items per objective; thus each reading test consists of 36 items and each math test is comprised of 30 items.

The main purpose of the testing program is to determine whether or not each student has met statewide performance standards in each of the eleven basic skills areas. In addition, diagnostic information regarding each objective is to be provided to facilitate the planning of remedial instruction for those students who fall short of the statewide minimum performance. Due to the small number of items covering each objective, student performance in each objective is categorized only as *Adequate* or *Non-adequate*. Also, adequacy classification for each objective is to be based on the statewide standard set for the test of which the objective constitutes a component.

This study will describe two latent-trait approaches to adequacy classifications for the BSAP objectives. One procedure is based on the one-parameter logistic (Rasch) model; the other one relies on the two-parameter logistic (2PL) model. Both techniques will be applied to the 1981 BSAP tests and the results will be compared.

2. Overall Procedure for Objective Adequacy Classification

Consider a test of L items; each item is scored zero or one and had an item characteristic curve (icc) described by the function $P(\theta)$. This quantity $P(\theta)$ is the probability that an examinee with ability θ will answer the item correctly. Let the test score be the number of correct responses. Then the test characteristic curve (tcc) of the test is the *expected* number of correct responses that an examinee with ability θ will make on the test. It is given as

$$E_L(\theta) = \sum_{j=1}^{L} P_j(\theta)$$

where $P_j(\theta)$ is the icc of the j-th item. Let c be the passing (cutoff) score on the test; that is, c is the minimum number of correct responses that an examinee must have in order to pass the test. The corresponding cutoff value on the ability (θ) scale is the value θ_c which satisfies the equation $E_L(\theta_c) = c$. This value may be found by using an appropriate iteration procedure such as the Newton-Raphson technique.

Now let the L-item test be divided into m subtests of length $L_1, L_2, \ldots, L_m$. Each subtest measures one objective. Without loss of generality, let the first subtest consist of the first L_1 items. The tcc of this subtest is given as

$$E_1(\theta) = \sum_{j=1}^{L_1} P_j(\theta).$$

At the cutoff ability θ_c, the expected number of correct responses on the first subtest is $E_1(\theta_c)$. Let c_1 be the smallest integer which is larger than or equal to $E_1(\theta_c)$. Then c_1 may be taken as the passing score on the first subtest. By the same procedure, the expected number of correct responses $E_i(\theta_c)$, $i = 2, \ldots, m$ of the remaining subtests may be determined. For each subtest, then, the passing score c_i may be taken as the smallest integer which is equal to or larger than $E_i(\theta_c)$.

The procedure presented above rests upon two assumptions. First, all items in the test tap the same ability dimension; hence the subtests may differ only in terms of difficulty. In other words, any content variation among the objectives does not bring in any extra ability factor; the variation in content reflects only the difficulty level with which each objective is placed on the common ability dimension. Second, the cutoff ability set for the common ability dimension applies to the test as well as each subtest.

It may be noted that the sum of the expected numbers of correct responses $E_1(\theta_c) + E_2(\theta_c) + \ldots + E_m(\theta_c)$ is exactly the test passing score c. However, when each c_i is equal to the value of $E_i(\theta_i)$ *rounded upward* to the nearest integer, the sum $c_1 + c_2 + \ldots + c_m$ is in general higher than c. Thus, students who *barely* pass all the objectives may have total test scores substantially higher than the (minimum) test passing score. This indicates that the passing scores for the objectives computed this way may be somewhat more stringent than are needed.

Another way to set passing scores for the objectives is to round each $E_i(\theta_c)$ to its nearest integer r_i under the constraint that $r_1 + r_2 + \ldots + r_m = c$. Although this rounding-off procedure does not hold constant the cutoff ability for each objective, it does guarantee that students who barely pass the objectives will barely pass the test. In addition, a student who barely passes some objectives and barely misses the remaining ones will not meet the test passing score. In the remaining part of this chapter, the r_i's will be referred to as *constant-sum passing scores*.

3. Iterations for Cutoff Abilities

This section will describe the Newton-Raphson iteration process for determining the cutoff ability θ_c. All items are presumed to have been calibrated; hence item difficulty and, where appropriate, item discrimination are known.

In the context of the Rasch model, each item is characterized by its difficulty δ and its icc is given as

$$P(\theta) = \exp(\theta-\delta)/\left(1 + \exp(\theta-\delta)\right).$$

To solve the equation for the cutoff ability θ_c, let

$$Q(\theta) = 1 - P(\theta) = 1/\left(1 + \exp(\theta-\delta)\right).$$

In addition, let

$$F = \sum_{j=1}^{L} P_j(\theta) - c$$

and

$$G = \sum_{j=1}^{L} P_j(\theta)Q_j(\theta).$$

Then with θ_c as the current approximate cutoff ability the Newton-Raphson updated cutoff ability is $\theta_c - F/G$. When c is not a zero or perfect score, a good starting value for θ_c may be taken as $\log\left(c/(L-c)\right)$.

In the two-parameter logistic model, each item is described by its discrimination α (a scale index) and its difficulty β (a location index). The icc is given as

$$P(\theta) = \exp\left(\alpha(\theta-\beta)\right)/\{1 + \exp\left(\alpha(\theta-\beta)\right)\}.$$

To apply the Newton-Raphson procedure in solving the equation $E_L(\theta_c) = c$ for the cutoff ability θ_c, let

$$Q(\theta) = 1 - P(\theta) = 1/\{1 + \exp\left(\alpha(\theta-\beta)\right)\}.$$

In addition, let

$$F = \sum_{j=1}^{L} P_j(\theta)$$

and

$$G = \sum_{j=1}^{L} \alpha_j P_j(\theta)Q_j(\theta).$$

Then with θ_c as the current approximate cutoff ability, the updated value is $\theta_c - F/G$. As in the Rasch case, an initial value for θ_c may be taken as $\log\left(c/(L-c)\right)$.

4. Item Calibration via the Rasch and Two-Parameter Logistic Models

As described in Chapter 1, the Rasch model was chosen as the general framework for all technical work. The decision was made primarily because the Rasch model is the logistic model which is most consistent with the tradition of using the number of correct responses as test scores. For each test administered in 1981, all items were calibrated on a sample of approximately 2600 students using a version of BICAL3 available at the University of South Carolina. The mean difficulty of items in each test was (arbitrarily) set at zero; these items defined an ability scale which was held in common for all the subtests covering the objectives. Tables 14-18 report the results of the Rasch calibration process.

To set the ground for adequacy classifications based on the two-parameter logistic model, the LOGIST program was used to determine the discrimination and difficulty parameters for the items in each test. As in the Rasch model, the item parameters in each test automatically determine an ability scale; this scale is treated as the common ability scale underlying the responses to items in each objective. The results of the LOGIST runs are documented in Tables 14-18.

5. Adequacy Classification for BSAP Objectives

On the basis of the item parameters reported in Section 4 and of the statewide passing scores listed in Table 5 of Chapter 1, cutoff ability values (θ_c) were computed using the Rasch and the two-parameter logistic (2PL) models for each reading and math test. Each θ_c value was then held constant for all objectives which form the test. Based on the item parameters and the θ_c values, the expected numbers of correct responses $E_i(\theta_c)$ were subsequently determined for each objective. These values were first rounded upward to the *rounded-upward passing scores* c_i. Then they were rounded to the nearest integers under the constraint $r_1 + r_2 + \ldots + r_m = c$; these are the *constant-sum passing scores*.

TABLE 14

Rasch and 2PL Item Parameters
for Reading and Math, Grade 1

Reading				Math			
	Rasch	2PL			Rasch	2PL	
Name	δ	α	β	Name	δ	α	β
DW01	-1.365	1.870	-1.249	OP04	0.918	0.493	-1.267
DW04	-1.446	1.355	-1.441	OP08	0.475	1.075	-1.121
DW10	-1.704	2.000	-1.337	OP11	0.710	0.696	-1.193
DW14	-0.994	0.819	-1.582	OP12	0.972	0.430	-1.341
DW16	-1.682	1.939	-1.330	OP13	1.168	0.623	-0.848
DW20	-1.169	1.506	-1.262	OP21	0.938	0.699	-0.956
MI03	0.547	1.985	-0.536	CN05	-1.618	1.156	-2.251
MI07	0.260	1.027	-0.703	CN07	0.029	1.003	-1.434
MI09	-0.300	1.080	-1.015	CN14	-1.053	0.767	-2.462
MI10	-0.324	0.909	-1.094	CN16	-1.174	0.698	-2.695
MI19	-0.561	1.160	-1.123	CN19	0.409	0.339	-2.443
MI20	-0.450	1.504	-0.969	CN20	4.012	0.010	79.288
DE02	0.615	0.744	-0.546	GE03	-0.106	0.091	-11.967
DE09	0.976	0.620	-0.300	GE11	0.540	0.401	-1.949
DE12	0.824	1.048	-0.359	GE02	-2.711	1.211	-2.709
DE13	0.598	0.858	-0.535	GE18	-2.007	0.858	-2.913
DE19	0.580	1.241	-0.509	GE19	0.318	0.504	-1.901
DE20	0.959	0.859	-0.301	GE21	0.253	0.566	-1.810
AL02	1.199	0.549	-0.139	ME04	0.230	0.502	-1.971
AL03	1.103	0.692	-0.197	ME05	0.645	0.888	-1.065
AL08	0.461	0.811	-0.640	ME09	-0.284	0.437	-2.848
AL09	1.169	1.057	-0.159	ME11	1.241	0.350	-1.176
AL19	0.776	0.689	-0.422	ME12	-1.547	0.942	-2.459
AL20	1.645	0.519	0.309	ME21	-1.817	2.000	-1.886
RE04	-1.279	0.996	-1.670	PS06	-0.176	0.935	-1.598
RE07	-0.216	1.748	-0.842	PS18	-0.106	0.983	-1.515
RE08	-0.433	1.333	-1.015	PS19	0.406	1.186	-1.100
RE09	-0.665	1.362	-1.094	PS04	-0.258	1.196	-1.494
RE15	-0.292	0.844	-1.120	PS05	-0.454	1.456	-1.464
RE16	-0.068	0.861	-0.937	PS17	0.051	1.208	-1.311
IN02	0.139	0.307	-1.738				
IN05	0.376	0.548	-0.845				
IN08	0.262	1.066	-0.698				
IN10	-0.090	1.888	-0.778				
IN13	0.226	0.728	-0.825				
IN17	0.320	1.049	-0.685				

TABLE 15

Rasch and 2PL Item Parameters
for Reading and Math, Grade 2

Reading				Math			
	Rasch	2PL			Rasch	2PL	
Name	δ	α	β	Name	δ	α	β
DW10	-2.313	0.547	-3.501	OP05	0.772	1.347	-0.971
DW13	-2.953	0.383	-5.639	OP10	0.659	0.827	-1.291
DW02	-0.394	1.190	-1.089	OP11	0.582	1.531	-1.030
DW03	-1.578	2.000	-1.424	OP14	2.288	0.404	-0.004
DW15	0.571	0.411	-0.967	OP19	0.713	0.693	-1.387
DW20	0.482	0.764	-0.706	OP20	-0.054	0.722	-1.972
MI05	0.034	0.134	-4.418	CN18	1.970	0.184	-0.805
MI07	0.316	0.362	-1.391	CN15	-0.449	0.770	-2.186
MI08	0.848	0.481	-0.580	CN10	0.135	0.937	-1.560
MI10	1.066	0.606	-0.330	CN04	-0.437	1.467	-1.551
MI13	0.965	0.368	-0.568	CN06	0.495	0.829	-1.396
MI17	0.953	0.613	-0.420	CN20	-0.019	1.413	-1.353
DE05	-0.072	1.227	-0.918	GE04	0.251	0.811	-1.607
DE06	-0.098	1.352	-0.923	GE06	0.445	0.863	-1.414
DE07	-0.084	1.507	-0.880	GE08	-0.675	0.619	-2.738
DE08	0.249	1.206	-0.742	GE13	-1.507	0.912	-2.642
DE12	-0.250	1.277	-1.000	GE14	-1.799	1.538	-2.070
DE22	-0.777	1.159	-1.314	GE15	-0.490	0.848	-2.063
AL04	0.726	1.217	-0.473	ME03	0.797	0.426	-1.865
AL07	0.492	1.160	-0.615	ME09	0.795	0.667	-1.340
AL12	0.199	1.078	-0.799	ME11	-0.414	0.493	-3.043
AL13	0.064	1.641	-0.796	ME13	-1.667	0.897	-2.789
AL15	0.721	0.770	-0.531	ME15	-2.147	0.502	-4.919
AL18	0.726	0.777	-0.523	ME21	1.696	0.380	-0.814
RE01	-0.716	1.091	-1.312	PS19	-0.965	0.989	-2.184
RE02	-0.323	0.823	-1.246	PS04	-0.551	0.718	-2.391
RE05	0.352	0.700	-0.835	PS14	-0.408	1.003	-1.829
RE09	-0.298	1.223	-1.034	PS15	0.772	0.872	-1.201
RE14	0.124	1.005	-0.853	PS07	-0.337	0.841	-1.985
RE17	-0.150	1.332	-0.926	PS02	-0.403	0.803	-2.122
IN01	-0.164	1.987	-0.889				
IN04	0.132	1.459	-0.790				
IN07	-0.044	1.215	-0.917				
IN13	0.185	0.777	-0.915				
IN16	0.482	1.944	-0.648				
IN17	0.532	1.230	-0.597				

TABLE 16

Rasch and 2PL Item Parameters
for Reading and Math, Grade 3

Reading				Math			
	Rasch	2PL			Rasch	2PL	
Name	δ	α	β	Name	δ	α	β
DW08	-1.991	0.010	-166.350	OP02	-0.115	0.807	-1.233
DW14	-1.053	0.835	-1.803	OP07	0.765	1.245	-0.453
DW05	-0.452	0.988	-1.287	OP08	0.497	1.257	-0.631
DW07	0.610	0.815	-0.668	OP12	-0.016	0.591	-1.403
DW15	-0.638	0.935	-1.426	OP15	0.612	1.056	-0.573
DW16	-0.226	0.975	-1.138	OP20	1.179	0.632	-0.243
MI05	0.371	0.518	-1.128	CN16	-0.177	0.752	-1.326
MI11	0.463	0.764	-0.799	CN13	0.821	0.708	-0.498
MI13	0.845	0.766	-0.506	CN08	-0.658	0.644	-1.920
MI14	0.761	0.709	-0.613	CN01	0.377	1.053	-0.732
MI17	0.857	0.602	-0.594	CN05	1.064	0.441	-0.370
MI20	0.543	0.596	-0.877	CN20	-0.507	0.275	-3.604
DE05	-0.160	1.341	-0.977	GE01	0.041	0.188	-3.633
DE06	-0.885	1.941	-1.243	GE08	-0.138	0.223	-3.523
DE11	-0.188	0.792	-1.238	GE09	-1.265	0.638	-2.495
DE12	0.306	0.256	-2.150	GE18	-1.516	0.856	-2.203
DE18	-1.192	1.945	-1.346	GE19	-1.094	0.719	-2.144
DE20	-0.788	1.438	-1.276	GE20	0.415	0.010	-48.215
AL02	-0.646	1.476	-1.188	ME01	-1.003	1.030	-1.657
AL05	-0.700	1.426	-1.249	ME08	-0.766	0.494	-2.468
AL13	-0.130	1.226	-0.994	ME04	-0.006	0.227	-3.128
AL14	-0.692	1.745	-1.183	ME15	-1.791	0.670	-2.867
AL17	2.327	0.079	4.213	ME20	0.501	0.386	-1.213
AL19	1.797	0.010	6.265	ME21	1.793	0.348	0.683
RE05	0.810	0.685	-0.557	PS06	-0.473	0.954	-1.368
RE07	0.179	0.820	-0.972	PS11	-0.264	0.360	-2.449
RE10	-0.142	0.658	-1.376	PS12	0.388	0.493	-1.135
RE14	0.082	0.593	-1.262	PS13	0.175	0.589	-1.206
RE17	0.239	0.632	-1.073	PS17	0.471	0.996	-0.682
RE20	-0.801	1.360	-1.289	PS21	0.688	0.990	-0.534
IN04	-0.268	1.685	0.977				
IN08	0.234	0.984	-0.860				
IN09	0.506	0.814	-0.739				
IN13	0.130	1.282	-0.845				
IN14	-0.571	1.119	-1.290				
IN21	0.461	0.814	-0.769				

TABLE 17

Rasch and 2PL Item Parameters
for Reading and Math, Grade 6

Reading				Math			
	Rasch	2PL			Rasch	2PL	
Name	δ	α	β	Name	δ	α	β
DW07	-0.405	1.222	-0.903	OP01	-1.274	0.442	-1.888
DW09	-1.223	0.794	-1.676	OP04	0.023	0.944	-0.202
DW11	-2.268	1.660	-1.643	OP11	-0.721	0.720	-0.856
DW12	0.527	0.614	-0.480	OP16	-0.084	1.051	-0.623
DW17	-1.511	1.576	-1.354	OP18	-0.304	0.955	-0.431
DW18	0.197	0.912	-0.615	OP21	1.122	0.564	0.846
MI01	1.476	0.648	0.368	CN15	1.037	0.353	1.138
MI06	0.425	0.844	-0.471	CN13	0.675	0.609	0.376
MI11	0.617	0.507	-0.441	CN01	-0.570	0.831	-0.663
MI12	0.353	0.700	-0.562	CN04	-0.120	0.675	-0.353
MI17	0.579	0.656	-0.405	CN11	1.074	0.331	1.247
MI21	0.506	0.622	-0.473	CN19	-1.406	1.270	-1.065
DE01	-0.387	1.012	-0.959	GE01	-0.848	0.397	-1.452
DE05	-0.910	1.284	-1.149	GE08	-0.107	0.229	-0.772
DE06	-0.447	1.191	-0.924	GE11	-0.195	0.415	-0.568
DE08	-0.591	0.785	-1.231	GE14	0.665	0.562	0.418
DE15	-0.126	0.837	-0.860	GE19	1.045	0.806	0.625
DE17	0.091	0.789	-0.733	GE21	1.926	0.517	1.833
AL02	1.124	0.456	0.096	ME01	-0.607	0.671	-0.775
AL03	0.822	0.944	-0.196	ME10	-0.179	0.482	-0.483
AL11	1.036	0.432	-0.012	ME05	-0.011	0.522	-0.263
AL14	0.850	0.559	-0.181	ME14	0.734	0.545	0.488
AL17	1.330	0.451	0.307	ME19	-0.176	1.049	-0.326
AL18	1.097	0.736	0.032	ME20	0.965	0.754	0.595
RE04	-0.531	0.941	-1.094	PS07	-0.200	1.176	-0.331
RE07	-1.204	1.147	-1.368	PS10	-0.663	1.200	-0.622
RE08	-0.964	1.031	-1.293	PS01	-0.870	0.953	-0.823
RE14	-1.524	1.036	-1.636	PS04	-0.237	0.509	-0.521
RE17	-1.175	0.610	-1.969	PS12	-0.872	0.809	-0.899
RE19	-0.392	0.947	-0.979	PS19	0.143	0.531	-0.113
IN01	0.267	0.968	-0.553				
IN07	0.538	0.931	-0.382				
IN09	-0.209	1.220	-0.802				
IN14	0.785	1.038	-0.214				
IN15	0.924	0.791	-0.124				
IN21	0.323	1.218	-0.492				

TABLE 18

Rasch and 2PL Item Parameters
for Reading and Math, Grade 8

Reading				Math			
	Rasch	2PL			Rasch	2PL	
Name	δ	α	β	Name	δ	α	β
DW06	0.505	0.755	-0.399	OP05	-0.285	0.766	-0.005
DW07	0.370	0.408	-0.750	OP09	-0.479	0.615	-0.157
DW09	-1.271	1.232	-1.375	OP10	-0.471	0.567	-0.170
DW10	-1.523	1.706	-1.356	OP13	-0.623	0.319	-0.530
DW16	-0.280	1.973	-0.883	OP19	-0.316	1.091	-0.046
DW21	0.168	0.737	-0.667	OP21	-0.176	0.512	0.154
MI03	0.189	0.719	-0.651	CN02	0.782	0.620	1.040
MI04	-0.302	1.059	-0.864	CN04	1.621	0.708	1.650
MI05	0.349	0.334	-0.892	CN08	-0.135	0.893	0.097
MI13	0.000	0.496	-1.025	CN16	-0.144	0.499	0.188
MI17	0.293	0.468	-0.755	CN19	-1.065	0.826	-0.545
MI21	0.747	0.889	-0.186	CN21	0.638	0.922	0.748
DE01	-0.181	0.791	-0.907	GE01	0.281	0.453	0.711
DE03	0.293	0.790	-0.545	GE03	0.290	0.598	0.559
DE04	-0.275	1.020	-0.873	GE07	-0.434	0.546	-0.111
DE11	-0.092	0.797	-0.834	GE11	0.699	0.338	1.551
DE14	-0.750	1.021	-1.165	GE14	0.928	0.561	1.270
DE17	-0.788	0.744	-1.400	GE16	-0.392	0.598	-0.072
AL02	0.927	0.410	-0.059	ME06	0.791	0.016	30.624
AL04	-0.106	1.064	-0.741	ME03	-0.578	0.571	-0.282
AL10	0.027	1.053	-0.669	ME10	-0.762	0.665	-0.400
AL13	0.753	0.678	-0.214	ME12	-0.069	0.665	0.238
AL20	0.825	0.742	-0.149	ME18	0.657	0.620	0.936
AL21	0.948	0.845	-0.044	ME19	-0.595	0.724	-0.240
RE01	-0.710	1.160	-1.087	PS03	0.603	0.930	0.693
RE05	-0.167	0.968	-0.811	PS09	0.512	0.760	0.715
RE07	0.315	0.660	-0.584	PS08	-0.488	0.983	-0.160
RE17	-0.361	0.774	-1.053	PS11	0.237	0.265	1.063
RE18	-0.891	0.967	-1.292	PS17	-0.891	0.574	-0.592
RE21	-0.748	0.956	-1.193	PS20	-0.137	0.359	0.273
IN03	-0.205	0.656	-1.041				
IN06	0.244	0.661	-0.652				
IN08	0.319	1.231	-0.451				
IN10	0.325	0.770	-0.546				
IN13	0.687	0.572	-0.295				
IN20	0.364	0.808	-0.499				

The results of these computations are reported in Table 19 for the reading tests and Table 20 for the math tests. An asterisk (*) indicates a disagreement between the Rasch and 2PL passing scores. Among the 66 cases under study, there is complete agreement between the Rasch and 2PL rounded-upward passing scores in 58 cases and a one-point disagreement in the remaining eight situations. As for the constant-sum passing scores, the Rasch and 2PL models provide identical results in 54 cases and a one-point disagreement in 12 cases.

It may be noted that the rounding-upward process yields a passing score of six (the perfect score) on a number of objectives. This occurs mainly for the reading tests in grades 1 and 2. Taking the fallibility of test data into account, these (perfect) passing scores may be somewhat more demanding than is typically necessary, especially for very young students.

In summary, for the BSAP tests administered in 1981, the Rasch and 2PL models provide subtest (objective) passing scores which are identical in the majority (about 80% to 90%) of situations. Due to the fact that test scores are taken as number of correct responses, the passing scores must be integers and can be obtained either by rounding upward or by rounding off to the nearest integer under the constant-sum constraint. The constant-sum passing scores are less demanding than the rounded-upward passing scores; they are perhaps more amenable to acceptance by teachers and other school personnel who have to deal with the basic skills assessment program.

6. An Historical Note

The rounded-upward and constant-sum passing scores based on the Rasch model were reported to the staff of the Office of Research of the South Carolina Department of Education in a meeting in February, 1982. It was recommended by Huynh Huynh that the Rasch constant-sum procedure be used with the constraint that the passing score for each objective be at least three (half of the number of items in each

TABLE 19

Rasch and 2PL Expected Number of Correct Responses $E_m(\theta_c)$ at True Cutoff Abilities and Passing Scores for BSAP Reading Objectives

Grade	Cutoff Ability		Objective	$E_m(\theta_c)$		Rounded-upward Passing Scores		Constant-sum Passing Scores	
	Rasch	2PL		Rasch	2PL	Rasch	2PL	Rasch	2PL
1	.542	-.436	DW	5.23	4.82	6	5*	5	5
			MI	3.95	3.79	4	4	4	4
			DE	2.68	2.99	3	3	3	3
			AL	2.26	2.79	3	3	2	3*
			RE	4.38	4.09	5	5	4	4
			IN	3.50	3.52	4	4	4	3*
2	1.099	.030	DW	4.98	4.75	5	5	5	5
			MI	3.57	3.52	4	4	3	3
			DE	4.66	4.68	5	5	5	5
			AL	3.87	4.04	4	5*	4	4
			RE	4.65	4.48	5	5	5	4*
			IN	4.26	4.54	5	5	4	5*
3	1.076	.119	DW	4.92	4.71	5	5	5	5
			MI	3.64	3.82	4	4	4	4
			DE	4.89	4.96	5	5	5	5
			AL	3.88	4.36	4	5*	4	4
			RE	4.34	4.30	5	5	4	4
			IN	4.34	3.86	5	4*	4	4
6	.834	.060	DW	4.76	4.64	5	5	5	5
			MI	3.26	3.39	4	4	3	3
			DE	4.61	4.40	5	5	5	4*
			AL	2.69	3.08	3	4*	3	3
			RE	5.11	4.74	6	5*	5	5
			IN	3.57	3.76	4	4	3	4*
8	1.033	.416	DW	4.62	4.74	5	5	4	5*
			MI	4.14	4.02	5	5	4	4
			DE	4.71	4.57	5	5	5	4*
			AL	3.66	3.89	4	4	4	4
			RE	4.82	4.70	5	5	5	5
			IN	4.05	4.09	5	5	4	4

Note: * indicates disagreement.

TABLE 20

Rasch and 2PL Expected Number of Correct Responses $E_m(\theta_c)$ at True Cutoff Abilities and Passing Scores for BSAP Math Objectives

Grade	Cutoff Ability		Objective	$E_m(\theta_c)$		Rounded-upward Passing Scores		Constant-sum Passing Scores	
	Rasch	2PL		Rasch	2PL	Rasch	2PL	Rasch	2PL
1	1.963	.737	OP	4.49	4.61	5	5	5	5
			CN	4.70	4.76	5	5	5	5
			GE	5.35	5.05	6	6	5	5
			ME	5.17	5.06	6	6	5	5
			PS	5.30	5.52	6	6	5	5
2	1.924	.342	OP	4.39	4.68	5	5	4	5*
			CN	4.86	4.95	5	5	5	5
			GE	5.45	5.32	6	6	6	5*
			ME	4.94	4.80	5	5	5	5
			PS	5.37	5.25	6	6	5	5
3	1.156	.329	OP	3.93	4.32	4	5*	4	4
			CN	4.28	4.31	5	5	4	4
			GE	4.97	4.61	5	5	5	5
			ME	4.49	4.380	5	5	5	4*
			PS	4.33	4.384	5	5	4	5*
6	.292	.115	OP	3.67	3.66	4	4	4	4
			CN	3.19	3.33	4	4	3	3
			GE	2.86	2.92	3	3	3	3
			ME	3.25	3.24	4	4	3	3
			PS	4.03	3.85	5	4*	4	4
8	-.009	.287	OP	3.56	3.36	4	4	3	3
			CN	2.63	2.78	3	3	3	3
			GE	2.67	2.78	3	3	3	3
			ME	3.13	3.08	4	4	3	3
			PS	3.02	3.01	4	4	3	3

Note: * indicates disagreement.

objective). This condition insures that the passing score for each objective is sufficiently above the chance score that would be obtained by randomly guessing at the answers. Thus, the final passing scores for the BSAP objectives were obtained by rounding off the values $E_{\square}(\theta_c)$ to the nearest integers under the condition that the results summed up to the statewide passing score and that each one of them was at least three. Table 21 reports the passing scores for each BSAP objective for the 1981 test administration.

TABLE 21

Passing Raw Score for Adequacy Status in Each Objective--BSAP 1981

		Grade				
Subject	Objective	1	2	3	6	8
Reading	DW	5	5	5	5	4
	MI	4	3	4	3	4
	DE	3	5	5	5	5
	AL	3	4	4	3	4
	RE	4	5	4	5	5
	IN	3	4	4	3	4
Math	CN	5	5	4	3	3
	OP	5	4	4	4	3
	ME	5	5	5	3	3
	GE	5	6	5	3	3
	PS	5	5	4	4	3

CHAPTER 3

A MINIMAX APPROACH TO SETTING MULTIVARIATE PASSING SCORES FOR SUBTESTS WHEN THE PASSING SCORE OF THE TOTAL TEST IS KNOWN

1. Introduction

In Chapter 2 a comparison was made on the use of the Rasch and two-parameter logistic models in setting passing scores for each objective in the South Carolina BSAP. At each grade level, the BSAP reading test consists of six six-item subtests measuring the objectives of decoding and word meaning (DW), main idea (MI), details (DE), analysis of literature (AL), reference usage (RE), and inference (IN). For each BSAP math test there are five six-item subtests focusing on the objectives of operations (OP), concepts (CN), geometry (GE), measurement (ME), and problem solving (PS). With the passing scores for the (total) reading and math tests at various grade levels already determined (see Chapter 1), the problem was to determine the passing score for each objective. The simultaneous setting of passing scores would be consistent in some sense with the passing score for the total test of which each objective was a part.

When the Rasch and two-parameter logistic models are used, strong assumptions are made on the relationship between the patterns of item responses and the examinee's ability. Moreover, in their current forms, logistic models assume that all test items tap the same unidimensional trait or ability and that item responses are coded as zero or one.

In many testing situations some of these assumptions may not be fully justified or feasible. For example, it may not be easy to document on the basis of content that math objectives such as concepts (CN) and problem solving (PS) can be conceptualized as parts of a common trait. In addition, many testing situations require the giving of partial credits or the scoring of test items on a scale other than from zero to one. For these cases, each item

response cannot be coded as zero or one; hence they cannot be framed within a typical binary logistic model.

Where test data are available for a group of examinees, the translation of the overall passing of a test to each of its objectives (subtests) may be accomplished within a (pseudo) decision theoretic framework. The purpose of this chapter is to describe a minimax approach to setting simultaneous passing scores for subtests when the passing score for the entire test is known in advance. The approach will be illuminated via its application to the South Carolina 1981 BSAP tests and the minimax objective passing scores will be contrasted with those based on the Rasch model.

2. The Minimax Procedure for Setting Multivariate Passing Scores

Consider now a test for which the test score is represented by Y. The test is divided into k subtests with the subtest (objective) scores denoted as $x_1, x_2, \ldots, x_k$. Thus $Y = x_1 + x_2 + \ldots + x_k$. Let c be the known passing score on the entire test. The problem at hand is to determine, simultaneously, k passing scores $\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)$ for the subtests in such a way that these subtest passing scores are consistent in some sense with the overall passing score c.

A preliminary observation may be made. Since the subtest scores sum up to the total test score, it appears desirable to have the $\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)$ such that the sum $r_1 + r_2 + \ldots + r_k$ is exactly c. This will insure that any examinee who barely passes each of the objectives will barely pass the entire test. This constraint will be maintained throughout the remainder of this chapter.

To set the stage of the minimax framework, let $p_i(r_i)$ be the proportion of examinees who are classified in the same way by the entire test and by the i-th subtest. In other words, p_i focuses on examinees for whom $Y < c$ and $x_i < r_i$, or $Y \geq c$ and $x_i \geq r_i$. With P denoting the probability of a given type of occurrence, p_i may be written as

$$p_i(r_i) = P(Y < c, x_i < r_i) + P(Y \geq c, x_i \geq r_i).$$

For each set of simultaneous passing scores $\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)$ let $p_{min}(\underset{\sim}{r})$ be the minimum of the p_i values computed for the k subtests. In other words

$$p_{min}(\underset{\sim}{r}) = \min\, p_1(r_1), p_2(r_2), \ldots, p_k(r_k) \ .$$

Within the minimax framework, the *optimal* simultaneous passing scores $\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)$ for the subtests correspond to the vector $\underset{\sim}{r}^o = (r_1, r_2, \ldots, r_k)$ such that the minimum probability $p_{min}(\underset{\sim}{r}^o)$ is the largest among all the probabilities $p_{min}(\underset{\sim}{r})$ computed for all possible configurations of $\underset{\sim}{r}$. Thus the minimax approach seeks to maximize the minimum probability of consistent classification between the total test and each of its subtests. (This is actually equivalent to minimizing the maximum probability of inconsistent classification between the total test and each of its subtests.)

The minimax approach can be implemented in a variety of ways. When each subtest score can take only a limited number of different values and when the number of subtests is not large, one may look at the entire region of $\underset{\sim}{r} = (r_1, r_2, \ldots, r_k)$ in which $r_1 + r_2 + \ldots + r_k = c$, compute $p_{min}(\underset{\sim}{r})$ at each $\underset{\sim}{r}$, and then search for the point $\underset{\sim}{r}^o$ at which this probability is the largest. The search can be accomplished in a fairly straightforward manner with the availability of a high-speed computer.

3. Illustrations Based on the South Carolina Basic Skills Assessment Program

The statewide passing scores for the 1981 BSAP reading and math tests are listed in Table 5 of Chapter 1. At each grade level and for each of the tests of reading and math, students were classified in two groups. The Failing group consisted of students with scores smaller than the statewide passing score. The Passing group was comprised of examinees for whom test scores equaled or exceeded the overall passing score. For each objective, the frequency distributions of the Failing and Passing groups were compiled and reported in Table 22 for the reading tests and in Table 23 for the math tests.

TABLE 22

Frequency Distributions of Scores in Reading Objectives for the Failing and Passing Groups

Grade	Objective	Group	Frequency at score						
			0	1	2	3	4	5	6
1	DW	Failing	14	42	93	138	152	237	257
		Passing	0	1	0	12	38	268	1674
	MI	Failing	51	128	203	254	176	97	24
		Passing	1	5	41	93	229	435	1189
	DE	Failing	101	246	315	195	60	16	0
		Passing	11	72	166	243	298	380	823
	AL	Failing	123	292	295	168	41	14	0
		Passing	15	91	215	311	371	401	589
	RE	Failing	25	119	164	237	200	102	86
		Passing	1	10	23	58	155	314	1432
	IN	Failing	40	149	282	250	163	45	4
		Passing	0	10	53	142	320	557	911
2	DW	Failing	4	15	68	203	309	305	107
		Passing	0	0	0	15	116	470	1066
	MI	Failing	34	102	290	312	184	63	8
		Passing	4	21	93	183	318	507	541
	DE	Failing	39	124	240	246	189	128	45
		Passing	0	0	6	21	76	278	1286
	AL	Failing	98	230	294	221	120	39	9
		Passing	0	1	32	84	214	364	972
	RE	Failing	25	100	207	269	213	147	50
		Passing	0	0	4	31	119	395	1118
	IN	Failing	71	214	261	222	133	88	22
		Passing	1	0	6	32	129	391	1108
3	DW	Failing	2	29	102	213	262	191	69
		Passing	0	0	4	17	130	437	1271
	MI	Failing	58	158	229	216	140	53	14
		Passing	1	21	102	184	243	454	854
	DE	Failing	34	78	129	150	203	175	99
		Passing	0	0	2	13	89	483	1272
	AL	Failing	35	106	185	244	196	91	11
		Passing	0	3	11	67	550	679	549
	RE	Failing	22	104	184	233	197	98	30
		Passing	0	4	11	78	203	575	988
	IN	Failing	46	160	228	192	126	85	31
		Passing	0	3	7	53	158	467	1171

TABLE 22

Frequency Distributions of Scores in Reading Objectives for the Failing and Passing Groups
(continued)

Grade	Objective	Group	Frequency at score						
			0	1	2	3	4	5	6
6	DW	Failing	21	60	163	269	311	255	89
		Passing	0	0	0	21	143	443	911
	MI	Failing	101	273	374	266	117	31	6
		Passing	1	20	76	214	403	444	360
	DE	Failing	63	126	181	249	240	210	99
		Passing	0	1	13	44	124	453	883
	AL	Failing	161	354	355	213	71	13	1
		Passing	6	46	167	313	373	377	236
	RE	Failing	22	60	103	196	293	315	179
		Passing	0	0	2	11	74	355	1076
	IN	Failing	142	304	311	237	120	43	11
		Passing	3	7	50	143	268	390	657
8	DW	Failing	76	94	200	332	329	219	69
		Passing	0	1	6	38	186	442	668
	MI	Failing	88	191	309	326	242	137	26
		Passing	0	0	27	113	280	459	462
	DE	Failing	65	137	253	267	272	226	99
		Passing	0	0	1	30	159	401	750
	AL	Failing	149	276	398	295	160	38	3
		Passing	0	7	45	145	294	416	434
	RE	Failing	79	120	197	249	308	248	118
		Passing	1	0	6	28	108	375	823
	IN	Failing	119	242	331	290	233	87	17
		Passing	0	6	24	85	241	436	549

TABLE 23

Frequency Distributions of Scores in Math Objectives for the Failing and Passing Groups

Grade	Objective	Group	Frequency at score						
			0	1	2	3	4	5	6
1	OP	Failing	29	108	192	244	218	104	51
		Passing	0	0	10	69	206	530	1165
	CN	Failing	7	16	57	152	328	349	37
		Passing	0	0	0	21	211	1121	627
	GE	Failing	6	8	34	124	212	272	290
		Passing	0	0	2	19	114	322	1532
	ME	Failing	4	13	60	148	276	298	147
		Passing	0	0	0	10	100	501	1369
	PS	Failing	14	48	100	164	179	197	244
		Passing	0	0	1	11	41	266	1661
2	OP	Failing	16	65	177	175	183	141	31
		Passing	0	0	16	69	204	619	982
	CN	Failing	8	15	76	158	243	216	72
		Passing	0	0	4	19	121	623	1123
	GE	Failing	7	6	24	60	177	249	265
		Passing	0	0	0	4	48	292	1546
	ME	Failing	8	3	29	125	243	292	88
		Passing	0	0	0	17	169	604	1100
	PS	Failing	8	9	29	103	138	281	220
		Passing	0	0	0	7	66	310	1507
3	OP	Failing	69	182	232	240	164	90	33
		Passing	1	11	35	152	282	499	737
	CN	Failing	20	67	205	276	259	150	33
		Passing	0	1	16	75	301	640	684
	GE	Failing	5	15	41	137	262	373	177
		Passing	0	0	2	37	188	636	854
	ME	Failing	6	17	82	246	384	226	49
		Passing	0	0	3	69	341	720	584
	PS	Failing	30	114	176	237	257	149	47
		Passing	1	6	30	101	240	467	872

TABLE 23

Frequency Distributions of Scores in Math Objectives for the Failing and Passing Groups *(continued)*

Grade	Objective	Group	Frequency at score						
			0	1	2	3	4	5	6
6	OP	Failing	83	297	378	319	180	78	13
		Passing	0	11	51	148	292	418	418
	CN	Failing	86	314	444	344	123	35	2
		Passing	0	10	105	259	430	335	199
	GE	Failing	108	325	465	337	101	12	0
		Passing	4	39	215	367	322	228	163
	ME	Failing	152	355	384	284	128	42	3
		Passing	6	20	107	171	365	393	276
	PS	Failing	115	226	273	317	264	124	29
		Passing	0	4	18	93	227	505	491
8	OP	Failing	160	366	510	362	165	67	11
		Passing	2	16	53	132	198	317	301
	CN	Failing	368	549	468	190	57	9	0
		Passing	1	35	115	247	278	217	126
	GE	Failing	314	519	491	251	64	2	0
		Passing	2	45	144	248	269	212	99
	ME	Failing	190	436	494	360	126	33	2
		Passing	5	16	92	193	300	304	109
	PS	Failing	212	486	503	310	108	21	1
		Passing	4	26	105	220	257	231	176

These tables serve as base data for the computation of the probabilities $p_i(r_i)$ of consistent classification between the total test and its i-th objective subtest. Let N be the number of students who took the test, N_F be the number of Failing students for whom the scores are less than r_i on the i-th objective, and N_P be the number of Passing students for whom the scores are at least r_i on this objective. Then the probability $p_i(r_i)$ can be estimated by the quantity $(N_F + N_P)/N$.

It may be recalled that each objective is measured by a subtest of six items. Thus the range for each r_i consists of all integers extending from 0 to 6. The search for the optimal simultaneous passing scores $\underset{\sim}{r}^o$ was confined to the set of vectors $\underset{\sim}{r}$ at which the sum $r_1 + r_2 + \ldots + r_k$ was equal to the overall passing score.

Tables 24 and 25 report the optimal minimax simultaneous passing scores for the BSAP objectives in reading and math, the p_i values (reported in percents) computed at these optimal passing scores, and the corresponding Rasch-derived passing scores reported in Chapter 2. An asterisk (*) is placed at the objectives for which a discrepancy exists between the minimax and Rasch-derived passing scores.

Among the 55 situations under consideration, there is complete agreement between the minimax and Rasch-derived passing scores in 39 cases. As for each of the remaining 16 cases, a discrepancy of <u>one</u> unit separates the minimax passing score from the one derived from the Rasch model. There is no apparent relationship between these discrepancies and the extent to which items in the corresponding objectives fit the Rasch model.

4. Discussion and Conclusion

A minimax scheme has been described for the simultaneous determination of passing scores for subtests (objectives) when the passing score for the whole test is known. The subtest passing scores are set up in such a way that there is maximum agreement between the pass-fail classifications based on the objectives and those classifications based on the whole test.

TABLE 24

Minimax and Rasch-Derived Simultaneous Passing Scores for 1981 BSAP Reading Objectives

Grade	Objective	Minimax Passing Score	Minimax p_i (%)	Rasch-Derived Passing Score
1	DW	5	81.4	5
	MI	4	85.1	4
	DE	3	82.2	3
	AL*	3	81.4	2
	RE	4	83.6	4
	IN*	3	82.1	4
2	DW	5	79.7	5
	MI*	4	79.1	3
	DE*	6	84.1	5
	AL	4	89.4	4
	RE*	4	83.4	5
	IN*	3	82.4	4
3	DW	5	84.9	5
	MI	4	81.1	4
	DE	5	86.1	5
	AL	4	86.1	4
	RE*	5	84.5	4
	IN*	3	83.7	4
6	DW	5	81.1	5
	MI	3	80.8	3
	DE	5	81.7	5
	AL	3	80.8	3
	RE	5	78.4	5
	IN	3	82.5	3
8	DW*	5	80.5	4
	MI	4	79.5	4
	DE	5	80.6	5
	AL*	3	79.4	4
	RE	5	80.9	5
	IN	4	83.0	4

TABLE 25

Minimax and Rasch-derived Simultaneous Passing Scores for 1981 BSAP Math Objectives

Grade	Objective	Minimax Passing Score	Minimax p_i (%)	Rasch-Derived Passing Score
1	OP	5	85.0	5
	CN	5	78.9	5
	GE	5	76.3	5
	ME	5	81.0	5
	PS	5	83.1	5
2	OP*	5	82.8	4
	CN	5	83.9	5
	GE*	5	78.9	6
	ME	5	78.9	5
	PS	5	78.6	5
3	OP*	5	77.9	4
	CN*	5	78.9	4
	GE	5	71.5	5
	ME*	4	73.2	5
	PS*	3	73.3	4
6	OP	4	82.1	4
	CN	3	77.0	3
	GE	3	73.6	3
	ME	3	78.0	3
	PS	4	80.2	4
8	OP	3	74.6	3
	CN	3	84.7	3
	GE	3	80.9	3
	ME	3	76.2	3
	PS	3	78.4	3

As applied to the reading and math objective subtests of the 1981 South Carolina BSAP, the minimax procedure provides passing scores which are identical to those derived from the Rasch model in about 70 percent of the cases. For the remaining 30 percent of the cases the discrepancy between each minimax passing score and the Rasch-derived cutoff score is one unit on each of the six-item subtests. Thus for all practical purposes, the minimax procedure and the Rasch model provide essentially the same passing scores for subtests similar to those of the South Carolina BSAP.

This study clearly demonstrates that in the setting of passing scores for subtests the minimax procedure is a viable alternative to a procedure based on latent trait models such as the Rasch when both approaches are applicable. Unlike latent trait models, the minimax approach does not impose strict assumptions on the way in which examinees respond to the test items and is applicable to simple 0-1 or more complex scoring schemes. The minimax approach is population dependent in the sense that it requires the administration of the entire test to a group of examinees. Latent trait models, on the other hand, rely on very strong assumptions about the nature of the test responses and require binary scoring for the test items. As long as test items have been calibrated, latent trait models can be used to set passing scores for subtests without the administration of the entire test to a group of examinees.

In a large (statewide or districtwide) testing program where the psychometric characteristics of binary test items are known in advance and when tests are to be administered to a large group of examinees, it is recommended that the minimax procedure and a suitable latent-trait scheme be used side by side in establishing passing scores for subtests. If the resulting cutoff scores are essentially the same, either of the two sets of passing scores may be chosen as final cutoff scores for the objectives. If they differ, it seems worth the effort to look carefully at the data and to explore the nature of the relationships among the subtests.

CHAPTER 4

REPORTING TEST SCORES AS PERCENT OF CORRECT RESPONSES AND CONSTRUCTION OF UNIT ITEMS IN A POOL

1. Introduction

In the previous two chapters, ways to classify student achievement on each objective are described. The classification is binary; that is, achievement in each objective is assessed only as *Adequate* or *Non-adequate*. Thus, for each test, passing scores are set simultaneously on each objective (six in reading and five in math) so that these passing scores are consistent with the overall passing score on the test. The purpose of providing information on each objective is to pinpoint the weaknesses of students who do not meet the statewide minimum standard in reading or math.

In a number of situations, it may be informative to report the percent of correct responses in each objective. When only one test form is used across years, the proportion of correct responses may be determined by dividing the number of correct responses by the number of items in each objective (six for the South Carolina BSAP). However, due to factors such as test security, different forms may be needed for different test administrations. Due to differences in item content and/or difficulty, these forms are not strictly equivalent. In other words, the same raw score (or percent of correct responses) may not bear the same meaning across different test forms. Hence, if test scores are to be reported in terms of percent of correct responses, procedures must be developed to take into account variation across different (alternate) test forms.

2. Proportion of Correct Responses in the Item Pool

Rather than using the proportion of items a student answered correctly on an objective (subtest), it may be more meaningful to relate his/her responses to the pool of items from which the subtest for the objective was assembled. Thus, for patterns of student

responses, an estimate is made of the proportion of items in the pool (which define the objective) which would be answered correctly. In this way, all percents of correct responses are expressed in terms of the items forming the item pool; thus a given percent would share the same meaning across different (alternate) forms even if these forms are not strictly equivalent.

Formally, let the item pool for a given objective consist of M items with item characteristic curves $P_i(\theta)$, $i = 1,2,\ldots,M$. From this pool, L items are selected to form the (sub)test for the objective. Without loss in generality, let us assume that these L items are indexed by $i = 1,2,\ldots,L$. The test characteristic function for the item domain is

$$E_M(\theta) = \sum_{i=1}^{M} P_i(\theta);$$

for the subtest, this function takes the form

$$E_L(\theta) = \sum_{i=1}^{L} P_i(\theta).$$

For an examinee with x correct responses on the subtest, the equation $E_L(\theta_x) = x$ will yield his or her ability θ_x. At this ability, the expected number of correct responses in the item pool is $E_M(\theta_x)$; hence the expected proportion of correct responses in the pool is $E_M(\theta_x)/M$. For the special case of $x = 0$ or L, the abilities are $\theta_0 = -\infty$ and $\theta_L = +\infty$; hence the expected proportions of correct responses in the item pool are respectively 0 and 1.

This procedure requires *a priori* calibration of all items in the pool; this may be done via the Rasch framework or most other latent trait models.

3. Illustration Based on the Rasch Model

As an illustration, let us consider the DW objective of the reading test for grade 1. There are 21 items in the pool; their

Rasch difficulty levels are listed in Table 26. For the 1981 BSAP test administrations, DW items included were DW01, DW04, DW10, DW14, DW16, and DW20. At the raw DW test scores of 1, 2, 3, 4, and 5, the θ_x abilities are -3.024, -2.098, -1.394, -.689, and .238. The corresponding expected number of correct responses in the pool and their percents (listed in parentheses) are 3.56 (17%), 6.77 (32%), 9.88 (47%), 13.05 (62%), and 16.52 (79%). At the raw score of zero, the percent of correct responses is zero; the percent is 100 at the maximum raw score of 6.

TABLE 26

Item Pool for DW Objective, Reading, Grade One

Item Name	Rasch Difficulty	Item Name	Rasch Difficulty	Item Name	Rasch Difficulty
DW01	-1.365	DW08	-1.441	DW15	-1.922
DW02	-1.667	DW09	-1.503	DW16	-1.682
DW03	0.677	DW10	-1.704	DW17	-1.701
DW04	-1.446	DW11	-0.070	DW18	-2.595
DW05	-0.421	DW12	-2.595	DW19	-1.922
DW06	-0.924	DW13	-1.102	DW20	-1.169
DW07	0.243	DW14	-0.994	DW21	-0.686

4. Psychometric Characteristics of the Unit Item of an Item Pool

When objective-referenced test scores are reported as percent of correct responses via the use of an item pool, it may be meaningful to conceptualize this pool as consisting of 100 uniform items (or unit items); thus the objective is psychometrically divided into 100 homogeneous units, each measured by one unit item, and all unit items are psychometrically identical.

With the pool consisting of M items, each with item characteristic curve $P_i(\theta)$, the (pool) test characteristic function is given as

$$E_M(\theta) = \sum_{i=1}^{M} P_i(\theta).$$

Thus, within the context of latent trait models, the unit item defining the pool has the item characteristic curve given as

$$\pi(\theta) = E_M(\theta)/M = \sum_{i=1}^{M} P_i(\theta)/M.$$

Since each $P_i(\theta)$ is monotonically increasing, the function $\pi(\theta)$ is also monotonically increasing; however, $\pi(\theta)$ may not share the same functional form with each $P_i(\theta)$.

Let $F(\theta)$ represent the distribution of the ability for a given population of examinees. Then, for the i-th item, the proportion of examinees who answer the item correctly (p-value) is given by the integral

$$p_i = \int P_i(\theta)dF(\theta).$$

The p-value of the unit item is the integral $\int\pi(\theta)dF(\theta)$; thus it is equal to the average

$$\sum_{i=1}^{M} p_i/M.$$

In other words, when the latent trait model fits the data adequately, the traditional difficulty (p-value) of the unit item is simply the mean p-value of all the items in the pool.

When each $P_i(\theta)$ follows a Rasch or two-parameter logistic model, the function $\pi(\theta)$ is probably fairly close to a two-parameter logistic function. Let

$$\hat{\pi}(\theta) = \exp[\alpha(\theta - \beta)/\{1 + \exp[\alpha(\theta - \beta)]\}$$

so that

$$\hat{\pi}(\theta)/[1 - \phi(\theta)] = \exp[\alpha(\theta - \beta)]$$

or

$$\alpha(\theta - \beta) = \log\{\hat{\pi}(\theta)/[1 - \hat{\pi}(\theta)]\}.$$

Thus the item parameters α and β of the unit item may be determined by fitting a straight line to the function $y(\theta) = \log\{\pi(\theta)/[1 - \pi(\theta)]\}$ at the ability values θ_x, $x = 1,2,\ldots,M-1$. (It may be noted that at θ_x, $\pi(\theta_x) = x/M$.)

Applied to the DW item pool listed in Table 24, the ordinary least squar- method yields the parameters $\alpha = .875$ and $\beta = -1.181$ for the *unit item* of the pool. When the test scores $x = 1,2,\ldots,20$ on the DW pool are expressed in terms of percents of unit items answered correctly, the discrepancies between the actual percents $100\pi(\theta_x)$ and the percents fitted via the unit item $100\hat{\pi}(\theta_x)$ do not exceed 2 percent.

It may be observed that the use of the Rasch model presumes that all items in the pool share the same degree of discrimination. However, when the pool is to be represented by its unit item, this unit item may or may not share the same level of discrimination. For the data of Table 26, all items have a discrimination value of one; however, the unit item has the factor .874 as its discrimination.

5. Potential Use of the Unit Item

Besides offering a unique description of the item pool, the concept of the unit item may be useful when the test constructor wishes to replenish the item pool without substantial changes in the statistical characteristics of the pool. To accomplish this, a two-parameter logistic (rather than the Rasch) model may be used to represent the item characteristic function. Then potential new items to be added to the item pool are those which match closely the difficulty and discrimination of the unit item underlying the pool.

PART C

EXPLORING THE USE OF PATTERNS OF INCORRECT RESPONSES

CHAPTER 5

EXPLORING THE USE OF PATTERNS OF INCORRECT RESPONSES IN SCORE REPORTING VIA THE BOCK MULTINOMINAL LATENT TRAIT MODEL

1. Introduction

Over the years researchers have explored the possibility of using the patterns of incorrect responses in multiple-choice items to extract more information from the examinees' responses. Given the constraints of classroom management, tests designed for diagnostic purposes such as those used in the South Carolina Basic Skills Assessment Program (BSAP) are relatively short. For these situations, the use of the raw score (i.e., the number of correct responses) would result in a loss of test data if more information could be derived from the patterns of incorrect responses. If these patterns are related to the ability level of the students and if they are taken into account in the scoring process, the resulting test scores may reflect more faithfully the achievement of these students.

A variety of procedures which consider both the correct option and the various incorrect options have been proposed for the scoring of tests with multiple-choice items. These procedures fall in two broad categories, those employing weighted option scoring and those using latent trait models.

In the first category, a weight of one is assigned to the correct option and other appropriate weights are given to the incorrect options. The score is then the sum of the weights of the options selected by the examinee. For each incorrect option, the weight depends on the seriousness of the error associated with this option. The weights may be determined empirically via point-biserial correlations or Guttman weights (see, for example, Claudy, 1978). They may also be based on expert judgements (Davis and Fifer, 1959; Downey, 1979). Research on the effectiveness of these weighted option scoring procedures has produced mixed results. None of the procedures

seems to result in test scores which are consistently more reliable or more valid than the raw scores.

The second category of test scoring based on responses to each of the options employs various latent trait models. Models developed by Samejima (1969, 1972) are appropriate for the analysis of items in which the options reflect various degrees of correctness or acceptability. For the scoring of multiple-choice items in which the options are basically nominal, Bock (1972) proposed a model based on a latent trait formulation of multinominal data. Both Bock (1972) and Thissen (1976) provided illustrations based on real test data. By use of the information function they stipulated that considerable gains in test score accuracy could be accomplished for lower ability examinees. Test information as defined by a latent trait model, however, reveals only an internal characteristic of the test; that is, if the model describes the data adequately, the test information will mirror the accuracy of the estimates obtained for whatever latent trait underlies the item responses. Hence, test information does not address the issue of the validity of test scores derived from the model.

This chapter will focus on the practicability of using the Bock model in scoring tests with multiple-choice items. It also will address the validity issue regarding ability estimates derived from this model. The research work was conducted using data from the sixth grade BSAP tests of reading and math. Since the BSAP tests for this grade are diagnostic, the conclusions reached in this study would be restricted to this type of data.

2. Overall Description of the Bock Multinominal Latent Trait Model

Consider a test with L multiple-choice items. When test scoring uses the raw score (i.e., the number of correct responses), each item is scored as one if the correct or best option is chosen; otherwise, the item will be scored as zero. This scoring treats all

incorrect options as equal and no provision is made for the seriousness of the error associated with each incorrect option. The latent trait model most congruent with number-of-correct scoring is the Rasch model. This model asserts that on an item with difficulty δ, an examinee with ability θ will give a correct response with a probability of

$$P(x = 1|\beta,\delta) = \frac{e^{(\beta - \delta)}}{1 + e^{(\beta - \delta)}} .$$

In the Bock model, the probability of selecting each of the options is considered separately. The model presumes that at each level of ability, the options have different probabilities of attracting the examinee. Thus, by taking these differences into account, better estimates for the ability would be obtained.

Let m_j be the number of options for the j-th multiple-choice item. Let k_j be any of these options. Then the use of the Bock latent trait model presumes that the probability of selecting this option be expressed as

$$P_{jk_j}(\theta) = \frac{\exp\left(Z_{jk_j}(\theta)\right)}{\sum_{h=1}^{m_j} \exp\left(Z_{jh}(\theta)\right)}, \qquad (1)$$

where

$$Z_{jh}(\theta) = C_{jh} + a_{jh}\theta, \; h = 1,2,\ldots,k_j,\ldots,m_j$$

and c_{jk} and a_j are the two item parameters associated with the h-th option of the j-th item (see Thissen, 1976, p. 202).

It may be noted from equation (1) that the probabilities associated with all the options are expressed via the same functional form; hence it is not possible to determine the correct option for an item by inspecting the option probabilities.

As in most latent trait models, the item parameters are presumed to be invariant across all subjects; in other words, they are

characteristics of the items and not of the examinees. When data are available, these parameters may be estimated (and the items are said to have been calibrated). There are many ways to estimate item parameters; the most commonly used are based on the maximum likelihood procedure. In this procedure, the item parameters are determined in such a way that the observed data are most likely to have come from the probability model underlying the estimated parameters. Bock (1972) described two estimation methods which are referred to as the conditional and unconditional procedures. Via LOGOG, the conditional estimation procedure has been implemented by Kolakowski and Bock (1973). LOGOG is used in this study.

Once all items are calibrated, the LOGOG program can be used to estimate the ability of (new) examinees who are not in the calibration sample. (It is assumed, of course, that the item parameters previously obtained will be applicable to these examinees.)

3. The Two Purposes of this Study

This study explores the feasibility of using the Bock model to score tests consisting of a limited number of multiple-choice items. Two questions are raised. First, does it make any difference whether raw scores or the Bock ability estimates are used to classify students? In other words, how strong is the relationship between the raw scores and the Bock ability estimates for tests with a moderate or small number of items? Second, how do the Bock ability estimates (as compared to the raw scores) relate to an external criterion when the criterion is used to validate the test or to set the passing score on the test?

4. Data Base, Item Calibration, and Ability Estimation

The data base of this study consisted of sixth graders to whom the BSAP tests of reading (2677 students) and math (2681 students) were administered in the spring of 1981. Teachers were asked to make judgements regarding their overall achievement in the above academic

areas. Some descriptive statistics regarding these students may be found in Tables 1 and 2 of Chapter 1. In addition to the item responses and teacher judgements, other background information such as race is available. It may be recalled that the teacher judgements were solicited prior to the administration of the BSAP tests; hence they are independent of the test data. There were three categories of teacher judgements: *Non-adequate*, *Adequate*, and *Undecided*. These judgements were used in the setting of passing scores for each of the BSAP tests.

The data base was then split into two parts via systematic sampling. The first part (one-third of the entire sample) was used to calibrate the items and the second part (two-thirds of the entire sample) served as the data base for the two research questions raised in the previous section.

The calibration was performed on the responses of 873 students for the reading test and of 892 for the math test. For the LOGOG program to run, the number of examinees choosing each option on each item could not be too small. Considering this constraint for each item, several options with low frequencies were combined in such a way that the total frequency would be at least 10 percent of the number of examinees in the calibration sample. This process seemed rather artificial since different options reflected different types of errors and usually the combined options did not share any other commonality than having low frequencies. However, this artificiality was the price to pay for convergence in the LOGOG program.

LOGOG required two passings, a diagnostic run and a final run. In the first run, examinees were sorted into 10 groups (fractiles) on the basis of the raw scores and initial estimates for the item parameters were obtained. These estimates were then used as starting values for the final run in which examinees were sorted into 10 fractiles on the basis of the estimated abilities.

Tables 27 and 28 present the data which reveal the degree to which the Bock model adequately describes the observed test data in

TABLE 27

Chi Square Tests of Goodness-of-Fit for Bock Abilities in Sixth Grade Reading

Item Sequence Number	Chi Square	Degrees of Freedom	Probability
01	21.6	16	>.05
02	8.4	8	>.05
03	17.5	8	<.05
04	32.9	24	>.05
05	6.0	8	>.05
06	8.3	16	>.05
07	26.1	24	>.05
08	47.7	24	<.01
09	40.5	24	<.05
10	43.0	24	<.01
11	25.4	24	>.05
12	19.0	24	>.05
13	38.9	24	<.05
14	445.3	16	<.01
15	16.3	24	>.05
16	56.1	24	<.01
17	45.7	24	<.01
18	20.0	16	>.05
19	30.4	24	>.05
20	28.8	24	>.05
21	35.0	24	>.05
22	51.0	24	<.01
23	41.3	24	<.05
24	36.2	24	>.05
25	30.5	24	>.05
26	23.4	16	>.05
27	20.2	16	>.05
28	30.5	16	<.05
29	8.2	8	>.05
30	20.5	24	>.05
31	23.6	24	>.05
32	40.6	24	<.05
33	22.9	24	>.05
34	24.7	24	>.05
35	34.7	24	>.05
36	31.0	24	>.05

TABLE 28

Chi Square Tests of Goodness-of-Fit for Bock Abilities in Sixth Grade Math

Item Sequence Number	Chi Square	Degrees of Freedom	Probability
01	19.0	16	>.05
02	19.8	24	>.05
03	25.5	24	>.05
04	16.3	24	>.05
05	30.5	24	>.05
06	21.0	24	>.05
07	25.7	16	>.05
08	24.0	24	>.05
09	16.9	24	>.05
10	21.4	24	>.05
11	45.6	24	<.01
12	28.9	16	<.05
13	26.2	24	>.05
14	33.3	24	>.05
15	18.3	16	>.05
16	17.6	16	>.05
17	19.5	16	>.05
18	35.2	16	<.01
19	22.9	24	>.05
20	34.5	16	<.01
21	24.6	16	>.05
22	37.3	24	<.05
23	28.2	24	>.05
24	28.1	24	>.05
25	27.6	16	>.05
26	34.5	16	<.01
27	30.9	16	<.05
28	33.0	24	>.05
29	47.6	24	<.01
30	26.2	24	>.05

the calibration sample. A small chi-square statistic indicates good fit whereas a large chi-square raises doubt about the appropriateness of the model. For the reading test probably six items do not fit the model whereas for the math test there are five such items (probability less than .01). Since this study focuses on the feasibility of using the Bock model for test scoring, there are no compelling reasons to delete items from the test.

With all items calibrated, LOGOG was then used to compute the ability estimates for the examinees not used in the calibration process. There were 1728 examinees for the reading test and 1764 for the math test. For each test, ability estimates were obtained for the entire test and for each of the subtests covering the objectives. (There were six objectives in reading and five objectives in math.)

Perfect or nearly perfect responses were observed for many examinees, particularly at the objective level. For these cases, successive LOGOG iterations resulted in estimates which drifted toward either $-\infty$ or $+\infty$. LOGOG then assigned the dummy estimates of -31 and 31 to these two nonconvergent cases.

To bring the nonconvergent estimates of -31 and 31 in line with the main body of the ability distribution, the minimum and maximum ability estimates for the convergent cases were determined for the entire test and for each of the subtests. For each case the smallest ability estimate was substituted for the dummy value of -31 and the dummy value of 31 was replaced by the largest ability estimate. Although this replacement of the nonconvergent values of -31 and 31 had no substantial statistical justification, it was done so that all examinees with perfect or near-perfect raw scores and those with zero or near-zero raw scores would be studied simultaneously with examinees in the middle of the raw score range. To delete cases with the nonconvergent values of -31 and 31 from the data analysis would grossly distort the testing framework within which the BSAP tests were assumed to function. In addition, this study focuses only on agreement between decisions based on the raw scores and those based

on the Bock abilities. In this context all conclusions will remain the same as long as the dummy ability -31 is replaced by any value smaller than the cutoff ability and the dummy ability 31 is replaced by any value larger than the cutoff ability.

5. Agreement Between Decisions Based on Raw Scores and Bock Ability Estimates

As stipulated, the Bock model is able to tap more information from the item responses than the raw scores at the lower end of the ability continuum. If this is the case, pass/fail classifications based on the Bock ability estimates would relate to those based on the raw scores to a lesser degree for students at the lower end of the ability scale than for those at the upper end.

The above assertion, however, could not be verified directly in this type of empirical study based on real data since the true ability of each student was not known. The assertion may be verified partially by noting that most students in this study had been classified by the teachers in one of three overall achievement categories (*Adequate*, *Non-adequate*, and *Undecided*). Though these classifications were made independently of the test data, they were strongly related to the test scores (see Chapter 1). Hence they may be used to sort students into groups which differ in overall ability. These groups would then be used to assess the differential relationship between raw scores and Bock ability estimates among groups with varying levels of ability stipulated in the previous section.

As may be recalled from Chapter 1, the passing score for each BSAP test was the median score of students for whom the teacher judgements were recorded as *Undecided*. For the reading and math tests used in this chapter, the passing scores are 24 and 17, respectively, on the raw score scale. For each test, students were placed in the passing group or the failing group based on these passing scores. As indicated in Chapter 2, pass/fail classifications were also made on the subtests covering the individual objectives. This

was done by translating the overall passing score into a cutoff score on a suitable Rasch ability scale. This cutoff ability was then used to compute the expected numbers of correct responses on the subtests. The results were rounded to the nearest integers in such a way that their sum equalled the overall passing score; these integers were finally used as cutoff scores for the objectives.

To set the framework by which pass/fail decisions could be made on the basis of the Bock ability estimates, the median ability of the *Undecided* group was used as the cutoff ability for the test under study. This cutoff ability was used to make pass/fail decisions on the entire test as well as on each of the objectives. (This process was not used strictly on the raw score scale because of the limited number of Rasch ability estimates for the objectives.)

On each of the reading and math tests and on each of the objectives, a pass/fail classification was made for each student on the basis of the raw score and another pass/fail classification was made using the Bock ability estimate. Agreement occurred if these two classifications produced the same result for the student; in other words, agreement occurred if the student was classified in the passing group by both the raw score and the Bock ability estimate or if the student was classified in the failing group by both these quantities. The proportion of students for whom these decisions are in agreement is typically referred to as an agreement index. In Figure 1 the agreement index is the proportion of students in the pass/pass and fail/fail categories.

For all students not included in the calibration subsample, an agreement index was computed for the reading and math tests and for each of the objectives. Agreement indices were also computed for students in the *Adequate* and *Non-adequate* samples. The results are compiled in Table 29.

The data clearly indicate that for the entire tests of reading and math, pass/fail classifications based on the raw scores and those

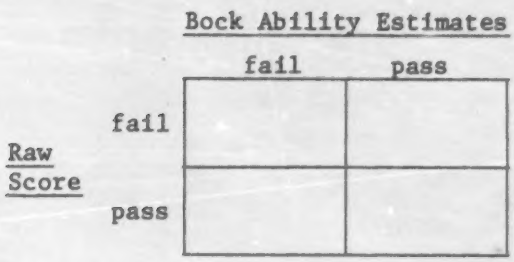

Figure 1. Decision consistency of pass/fail classifications based on raw score with classifications based on Bock ability estimates.

TABLE 29

Percent of Students Consistently Classified in the Same Categories by the Raw Score and Bock Ability

Reading				Math			
	Percent Consistency				Percent Consistency		
Test	Total	Non-adequate	Adequate	Test	Total	Non-adequate	Adequate
Entire test	95.5	95.6	96.2	Entire test	94.2	94.5	94.2
DW	95.0	93.2	96.8	OP	81.3	77.1	85.0
MI	88.0	84.0	91.3	CN	79.2	74.8	82.0
DE	79.7	71.0	85.8	GE	63.6	66.5	61.9
AL	80.4	72.9	86.2	ME	82.0	80.5	83.2
RE	83.7	74.3	90.7	PS	89.6	87.5	91.8
IN	89.7	82.6	94.5				

based on the Bock ability estimates are almost identical for all students under consideration as well as for those in the adequate and non-adequate subsamples. No differences seem apparent between these two groups at the entire test level.

At the objective level, less agreement was observed for all cases except the GE objective of the math test. For the six reading objectives, the agreement indices averaged 79.7 percent for the *Non-adequate* group and 90.9 percent for the *Adequate* group. For the four math objectives (GE excluded), these averages were 77.3 percent and 80.8 percent respectively.

The data appear to indicate that as long as the test has moderate length the Bock model and the use of raw scores produce almost identical pass/fail decisions even for students at the lower end of the ability continuum. However, when the test is short, pass/fail classifications based on the Bock ability estimates and those based on the raw scores appear to show less agreement for students at the lower end than for those at the upper end.

6. Relationship Between Pass/Fail Classifications and Teacher Judgements

In order to shed light on the validity of the pass/fail decisions based on the Bock model compared with the validity of those based on the raw scores, the teacher judgements (*Adequate* or *Non-adequate*) were used as an external validity criterion. There appeared to be no logical defense for the use of this criterion except that a teacher who had been teaching a student for almost nine months should be in a position to make a summative judgement regarding the overall achievement of the student. (No attempt was made to assess the reliability of the teacher judgement.)

For each of the reading and math tests and for each of their objectives, a four-corner table (Figure 2) was set up to record the number of students classified as pass or fail by the raw score (or Bock ability estimate) and as *Non-adequate* or *Adequate* by the teacher

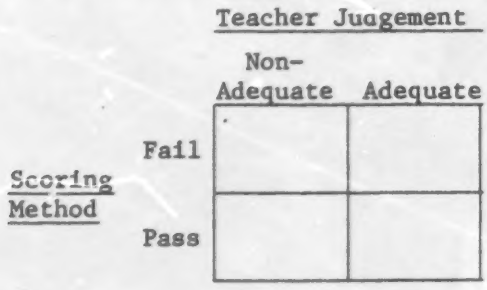

Figure 2. Decision consistency of teacher judgements with pass/fail classifications based on scoring methods.

judgement. Agreement occurred if the student was placed in the corner "pass/adequate" or "fail/nonadequate."

Table 30 reports the agreement index between pass/fail decisions and teacher judgements. The data clearly indicate that for both the entire tests of reading and math, there is no noticeable difference between the "validity" of pass/fail decisions based on raw scores and the validity of those decisions based on the Bock ability estimates. (For the reading test the agreement index is about 80 percent and for the math this index is near 75 percent.) Validity, of course, was judged by using the teacher judgements.

TABLE 30

Percent of Students Classified in the Same Categories by Teacher Judgements and Scoring Methods

Reading			Math		
	Percent Agreement			Percent Agreement	
Test	Raw Score	Bock Ability	Test	Raw Score	Bock Ability
Entire test	79.4	79.6	Entire test	74.5	74.5
DW	74.3	73.4	OP	71.1	69.8
MI	72.3	70.9	CN	66.5	62.8
DE	73.3	69.8	GE	65.4	57.1
AL	72.5	69.4	ME	68.4	65.5
RE	71.9	66.7	PS	72.1	72.1
IN	74.6	70.5			

Using the same criterion of validity, the picture changed considerably for pass/fail decisions based on each objective. The use of the Bock model resulted in pass/fail decisions less related to the teacher judgement than those decisions based on the raw score. For the six objectives of the reading test, the agreement index averaged 73.2 percent for the raw scores and 70.1 percent for the Bock ability estimates. For the five objectives of the math test, these averages were 68.7 percent and 65.5 percent, respectively.

Under the situations considered in this study, it appears that the use of the Bock model for a test with moderate length does not

change the validity of the pass/fail decisions in any noticeable way. On the other hand, if an external criterion such as teacher judgement is acceptable, the Bock model applied to a short test may result in pass/fail decisions which are less valid than those based on raw scores.

7. Concluding Remarks

This study indicates that in the context of pass/fail decisions, the use of the Bock multinominal latent trait model for moderate-length tests does not produce decisions which differ substantially from those based on the raw scores. Nor does the Bock model provide pass/fail decisions which are more valid than those based on raw scores when an external criterion such a teacher judgement is used.

On the other hand, for very short tests the pass/fail decisions based on the Bock model may differ somewhat from those decisions based on the raw scores. Thus, for very short tests, the ability tapped by the Bock model appears to differ from the one implied by the raw scores. Moreover, the Bock pass/fail decisions appear to relate less strongly to an outside criterion such as teacher judgement than those based on the raw scores. This anomaly makes it difficult to interpret the nature of the trait that the Bock model attempts to recover from the student responses.

This study demonstrates that when test data are used to make pass/fail decisions on students, the Bock model does not result in any differences from the use of raw scores when the test is of moderate length. Considering the complexity in item calibration and ability computations, the use of the Bock model does not seem to be justified. When the test is short, the Bock model appears to reflect a trait which is in variance with the one measured by the raw scores and reflected in an external criterion such as teacher judgement. This makes it difficult to interpret the nature of the trait revealed by the Bock model for these short tests. Thus, for these situations too, the Bock model does not appear to be useful.

Perhaps the Bock model may not be suitable for use with achievement items where a correct response exists. The functional form of the probability that the Bock model assigns to each option does not reveal any asymmetry regarding the correct and incorrect options; perhaps this lack of asymmetry accounts for the lack of positive results encountered. On the other hand, teacher judgements may not have been a good external criterion for the judgement of the validity of the trait implied by the Bock model. However, if the Bock model provided better estimates for ability than other estimates based on raw scores, this conclusion would have been tested against an acceptable criterion which is independent of the Bock estimates.

References

Bock, R. D. Estimation item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Claudy, J. G. Biserial weights: A new approach to test item option weighting. Applied Psychological Measurement, 1978, 1, 25-30.

Davis, F. B., and Fifer, G. The effect of test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 2, 159-170.

Downey, R. G. Item-option weighting of achievement tests: Comparative study of methods. Applied Psychological Measurement, 1979, 3, 453-461.

Kolakowski, D., and Bock, R. D. LOGOG: Maximum likelihood item analysis and test scoring: Logistic model for item responses. Chicago: National Educational Resources, 1973.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika, Monograph Supplement, No. 17, 1969.

Samejima, F. A general model for free-response data. Psychometrika, Monograph Supplement, No. 18, 1972.

Thissen, D. M. Information in wrong responses to the Raven Progressive Matrices. Journal of Educational Measurement, 1976, 13, 201-214.

CHAPTER 6

EXPLORING THE USE OF THE LOG-LINEAR MODEL IN THE IDENTIFICATION OF GROUP DIFFERENCES IN PATTERNS OF INCORRECT RESPONSES

1. Introduction

A major function of diagnostic testing and basic skills assessment is to identify weaknesses of students for suitable remediation. Typically, weaknesses are revealed through low scores; a diagnostic profile for a student can be composed if there are enough items to cover most major types of errors which need to be corrected. Most basic skills tests such as those used in the South Carolina Basic Skills Assessment Program (BSAP) are relatively short; therefore the use of raw scores (number of correct responses) does not permit a detailed analysis of student deficiencies.

An analysis of the patterns of incorrect responses may be helpful in mapping remediation strategies for students who need help. Due to the small number of items in most basic skills tests, such analysis may not be suitable for each individual student. However, if patterns of incorrect responses are related to identifiable student characteristics such as overall achievement, ethnicity, sex, or parental socioeconomic status, then students may be grouped on the basis of these characteristics in such a way that each group displays a different pattern of incorrect responses. If this type of analysis is appropriate, then a common remediation strategy can be adopted for each group of students.

The search for patterns of incorrect responses among subgroups of students may be of practical value to local schools or school districts which, due to limited financial resources, cannot devise individual remedial programs for all students who need help. A feasible way would be to group students on relevant characteristics (associated with the patterns of incorrect responses) and then to provide for each group a common strategy for rectifying the errors encountered in the acquisition of the subject area.

If students are grouped by student characteristics for remediation purposes, then these characteristics must display a substantial level of interaction with the errors made by students. For errors which are used as distractors in multiple-choice items, the selection of these student characteristics may be accomplished via an appropriate application of the log-linear model.

The purpose of this chapter is to provide illustrations of the application of the log-linear model to the selection of student characteristics which are relevant to the differential patterns of incorrect responses in multiple-choice items. The illustrations are based on responses of a large sample of sixth graders who took the BSAP reading test in 1981.

2. The Log-Linear Model in the Context of Analysis of Patterns of Errors

Consider a multiple-choice item with each distractor reflecting a different type of error. Let E be the variable representing these errors. (Hence each value of E corresponds to one distractor or one type of error.) Let $k = 1,\ldots,K$ be the index which ranges over the values of E.

As an illustration, let the student characteristics be denoted as A (with a different values) and B (with b different values). Let the i and j be the indices (subscripts) associated with A and B.

Within this context, the incorrect responses on the multiple-choice item may be sorted in a three-way A × B × E contingency table. Let f_{ijk} be the observed frequency (number of students) in each (i, j, k) cell of the table. Let F_{ijk} be the expected frequency of this cell. Under the log-linear model, $\ln F_{ijk}$ is the sum of several parameters. In the *full model*, a large number of effects due to the factors A, B, and E and their interactions are considered. For this case, $\ln F_{ijk}$ takes the form

$$\ln F_{ijk} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{E}_{k} + \lambda^{AB}_{ij} + \lambda^{AE}_{ik} + \lambda^{BE}_{jk} + \lambda^{ABE}_{ijk}.$$

As in the case of traditional analysis of variance, linear constraints are imposed on the parameters λ. They are

$$\Sigma\lambda^{A}_{i} = \Sigma\lambda^{B}_{j} = \Sigma\lambda^{E}_{k} = 0,$$

$$\Sigma_{i}\lambda^{AB}_{ij} = \Sigma_{j}\lambda^{AB}_{ij} = \ldots = \Sigma_{k}\lambda^{BE}_{jk} = 0,$$

and

$$\Sigma_{i}\lambda^{ABE}_{ijk} = \Sigma_{j}\lambda^{ABE}_{ijk} = \Sigma_{k}\lambda^{ABE}_{ijk} = 0.$$

The likelihood ratio statistic associated with this model is

$$G^2_F = 2 \sum_{i,j,k} f_{ijk} \ln (f_{ijk}/F_{ijk})$$

which is distributed asymptotically as chi-square with $n - p_F$ degrees of freedom (df) where n is the number of cells and p_F is the number of estimated independent parameters. For the present situation, $n = Kab$.

It may be noted that when the effects due to *all* the factors A, B, and E and to *all* their interactions are parts of the full model, the log-linear model provides complete fit to the data. For such a case, G^2_F and its df are zero.

If there are logical or practical reasons to consider factor A as the major variable in classifying students for the purpose of analysis of error patterns, then the interaction AE should be more substantial than the two combined interactions BE and ABE. Thus, the following *restricted model* may be used to describe the data

$$\ln F_{ijk} = \theta + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{E}_{k} + \lambda^{AB}_{ij} + \lambda^{AE}_{ik}.$$

Under this model, the chi-square statistic is given as

$$G^2_R = 2 \sum_{i,j,k} f_{ijk} \ln (f_{ijk}/F_{ijk})$$

which is distributed asymptotically as chi-square with $n - p_R$ degrees of freedom. Here, p_R is the number of independent parameters to be estimated under the restricted model.

It follows from the previous consideration that, after partialling out the interaction AE (i.e., the contribution of factor A to the explanation of the error variable E), the additional contribution of factor B to the explanation of the error variable E is given as

$$G^2_{add} = G^2_R - G^2_F.$$

Under the null hypothesis of no additional contribution in the population, G^2_{add} is a chi-square with $p_F - p_R$ degrees of freedom (see Bishop, Fienberg, & Holland, 1974, Section 14.9.6).

When the full model includes all the factors A, B, E and their interactions, $G^2_F = 0$ and $p_F = n$. In this case, the additional contribution of factor B to the explanation of the error variable E is G^2_R which is the chi-square with df = $n - p_R$. For the illustration purposes of this paper, this full model was used. All computations were carried out via the BMD computer program P4F (Dixon & Brown, 1981).

3. First Illustration: The Knowledge of Race Provides No Additional Information on Item 32

As the first illustration of the log-linear model to analysis of patterns of errors, two factors were used to group students: overall achievement (A) and race (B). Overall achievement had two categories, *Adequate* and *Non-adequate*. For race, the two categories were *Black* and *White*.

The data base consisted of 2252 sixth-graders who responded to the 32nd item of the BSAP reading test. They were students in the sample used in the setting of passing score for the sixth-grade reading test (see Chapter 1). This item had three incorrect options. Prior to test administration, judgements on student overall achievement were solicited from teachers who have taught the students in the sample. There were three categories of judgement, *Adequate, Non-adequate,* and *Undecided*. The *Undecided* category was very small; it was deleted in the analysis of patterns of errors.

Thus for the situation under consideration, the numbers of categories are a = 2 for factor A, b = 2 for factor B, and k = 3 for Factor E (which represents the three incorrect options). A total of 490 students did not respond to the item correctly; their frequencies in the 2 × 2 × 3 cells of the A × B × E contingency table are reported in Table 31.

TABLE 31

Frequency of Responses for Item 32

		E		
A	B	(1)	(2)	(3)
Ready	Black	15	27	77
	White	28	54	143
Non-adequate	Black	73	130	144
	White	49	72	126

Traditional contingency analyses on the marginal tables yielded the chi-square statistics of 26.23 (df = 2, $p < .01$) for the A × E table, 11.18 (df = 2, $p < .01$) for the B × E table, and 43.26 (df = 1, $p < .01$) for the A × B table. These analyses suggest a substantial association between the two factors A (overall achievement) and B (race). (The strength of the association may have been the result of factors including the cumulative effect of access to educational opportunity and cumulative effect of generations of social neglect on the part of black students.) With the two factors A and B highly correlated, any level of association between the factors A and E would also be reflected between the factors B and E and vice versa. Hence separate contingency analyses on the tables A × E and B × E would provide results which are highly dependent upon each other.

A multiple contingency analysis for the table A × B × E would be most meaningful since it provides a simultaneous consideration of the effects of the factors A and B on the patterns of errors represented by E.

Table 32 reports the results of the log-linear analyses for the data of Table 31 via seven models. Each of the models 2 through 7 contains the interaction of the error variable E with either A or B or both A and B. Of the two models which fit the data reasonably well (Model 4 and Model 7), Model 4 (G^2_R = 6.86, df = 4, p = .14) is the one which describes the data with the smaller number of terms.

TABLE 32

Results of Log-linear Fitting to Item 32

Model	Terms Included	df	Likelihood Ratio G^2	Probability
1	A, B, E, AB	6	33.09	.00
2	A, B, E, BE	5	65.16	.00
3	A, B, E, AE	5	50.12	.00
4	A, B, E, AB, AE	4	6.86	.14
5	A, B, E, AE, BE	3	38.94	.00
6	A, B, E, AB, BE	4	21.90	.00
7	A, B, E, AB, AE, BE	2	.70	.70

The data presented in Table 32 clearly indicate that, after the interaction between A (overall achievement) and B (race) has been partitioned out, the additional inclusion of the interaction AE in the model reduced the likelihood ratio G^2 from 33.09 to 6.86. This reduction of 26.23 is a chi-square with 6 - 4 = 2 degrees of freedom under the null hypothesis of no additional AE effects. Clearly, the effect due to AE is significant.

With both the interaction AB and AE in the model, the additional inclusion of BE reduced the G^2 from 6.86 to .70. This reduction of 6.16 (df = 2) is not significant.

In summary, the log-linear analyses presented above indicate that the association between race and the error variable can be traced to the relationship between overall achievement and the error variable. Hence for the item under study, the knowledge of the student's race did not appear to provide substantial information in explaining the pattern of incorrect responses displayed in the error variable.

4. Second Illustration: The Knowledge of Race Provides Additional Information on Item 4

To provide a second illustration on the use of the log-linear model in the analysis of error patterns, responses of sixth-graders to the BSAP reading item 4 were used. A total of 917 chose one of the three incorrect options; their frequencies are listed in Table 33.

TABLE 33

Frequency of Responses to Item 4

		E		
A	B	(1)	(2)	(3)
Ready	Black	53	20	51
	White	91	26	122
Non-adequate	Black	161	72	86
	White	106	34	95

Traditional chi-square analyses on the marginal tables resulted in the chi-square values of 48.80 ($df = 1$, $p < .01$) for the table A × B, 21.83 ($df = 2$, $p < .01$) for the table A × E, and 24.68 ($df = 2$, $p < .01$) for B × E.

The data of Table 34 indicate that among all the models under consideration, Model 7 is the only one which provides reasonable fit to the data. This model includes both interaction terms AE and BE; thus both factors A and B are needed to explain the variation in the types of errors students made on Item 4.

TABLE 34

Results of Log-linear Fitting to Item 4

Model	Terms Included	df	Likelihood Ratio G^2	Probability
1	A, B, E, AB	6	38.71	.00
2	A, B, E, BE	5	62.84	.00
3	A, B, E, AE	5	65.69	.00
4	A, B, E, AB, AE	4	16.89	.00
5	A, B, E, AE, BE	3	41.01	.00
6	A, B, E, AB, BE	4	14.03	.00
7	A, B, E, AB, AE, BE	2	.49	.78

5. Summary

This chapter focuses on the identification of student characteristics which relate substantially to the types of errors displayed in the distractors of multiple-choice items. The process may be implemented by selecting a number of relevant student characteristics, compiling the frequencies of students in the multiple contingency table defined by those characteristics and the types of errors, and finally by fitting an appropriate log-linear model to the table. Any student characteristic which interacts with the types of errors would be needed to account fully for these errors; hence they would be needed in the classification of students according to the type of errors.

This study focuses only on the methodology of selecting student characteristics which may be useful in describing the type of errors made on multiple-choice items. It does not address the nature of these types of errors. However, once a major student characteristic has been found to account for a substantial part of variation in types of errors made by students, one may take a look at these errors and see if they can be sorted into a small number of categories. The most common mistake made at each level of the said student characteristic may be reported; this information may be useful in the planning of instructional remediation.

References

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. Discrete Multivariate Analysis. Cambridge, MA: The MIT Press, 1974.

Dixon, W. J., and Brown, M. B., editors. BMDP Biomedical Computer Programs P-series. Berkeley, CA: University of California Press, 1981.

PART D

CONSIDERATIONS FOR BUDGET ALLOCATION

CHAPTER 7

ASSESSING THE BUDGETARY IMPACT OF REMEDIATION IN BASIC SKILLS ASSESSMENT PROGRAMS: I. STATISTICAL CONSIDERATIONS

1. Introduction

A major purpose of diagnostic testing and basic skills assessment programs is to identify students' strengths and weaknesses in certain academic areas so that remedial instruction can be given to students whose level of achievement is not up to par. Such areas typically include reading, writing, and mathematics. In general, for each basic skills area an overall test is given to assess the general level of achievement. Each student's overall score is compared to a predetermined passing score, with students scoring below the passing score being judged as non-adequate. For these students, the various subtests of the overall test are analyzed more thoroughly to identify the sub-areas which need attention.

The South Carolina Basic Skills Assessment Program (BSAP) exemplifies the use of test data for diagnostic purposes. Near the end of each school year, BSAP tests in reading and math are administered to students in grades one, two, three, six, eight, and eleven. (In addition, writing exercises are also given to students of grades six, eight, and eleven.) There are six objectives in reading: decoding and word meaning (DW), main idea (MI), details (DE), analysis of literature (AL), reference usage (RE), and inference (IN). In math, there are five objectives: operations (OP), concepts (CN), geometry (GE), measurement (ME), and problem solving (PS). Except for grade eleven, each objective is measured by a six-item subtest; thus each reading test consists of 36 items and each math test is comprised of 30 items. For grade eleven, each objective is covered by a subtest of 10 items.

At each grade level, a statewide passing score has been established for the reading and math tests. (See Chapter 1 for grades one, two, three, six, and eight.) By use of the Rasch constant-sum

procedure, the overall passing score was then translated into a passing score for each of the objectives assessed by the overall test. The translation was carried out in such a way that the individual objective passing scores are, in some sense, consistent with the overall passing score. Using these passing scores, each student's level of achievement on the overall test and on each objective can be assessed. Performance is said to be adequate for test scores at least equal to the passing score; otherwise it is deemed non-adequate.

For the South Carolina BSAP, the overall passing scores are used to identify students who might need additional instruction. Since the amount of remedial instruction depends on the number of objectives yet to be mastered, the cost of remediation varies from student to student. It would be ideal if remedial instruction could be provided to all students who need help, but the reality of budgetary constraints imposes a limit on the amount of additional instruction available. Thus, in setting passing scores in basic skills assessment programs, some concern should be given to the budgetary implications of choosing a particular cutoff score.

The issue of budgetary concerns in the setting of passing scores has been addressed by Huynh (1980). The general model provided in this study assumes that the cost of remediation can be assessed as a function of the true ability of the student. Given the remediation cost function $\delta(\theta)$ and the various probabilities associated with true ability and observed score, the budgetary consequences associated with a given cutoff score can be assessed. From the overall framework, details are presented for the special cases in which normal test scores follow either the beta-binomial model or the bivariate normal model.

The model provided by Huynh (1980) may be useful if remediation is given for the subject area covered by the overall test. In the context of basis skills assessment, however, remedial instruction is typically contemplated for the objective(s) or sub-area(s) in which the student appears to be weak. Therefore the previously mentioned

model needs to be extended to cover the case of basic skills assessment programs.

The purpose of this chapter is to provide ways to assess the budgetary implications of the various decisions regarding the setting of passing scores in basic skills assessment programs. The chapter also addresses the issue of equitable allocation to local school districts of funds designated for remedial instruction.

2. An Overall Framework

The chapter restricts the consideration of budgetary implications to situations in which the academic area covered by the overall test can be described by a unique latent trait. This restriction is consistent with the use of latent trait models such as the beta-binomial and the Rasch. (The beta-binomial model is a special case of the Rasch; it presumes that all test items are of equal difficulty.) As previously elaborated, the Rasch model has been used as the major vehicle for dealing with the several technical issues associated with the South Carolina BSAP.

Consider an academic area (such as math or reading) which is assessed via an overall test of n items. The area is divided into m sub-areas called objectives, each measured by a subtest of length $n_1,\ldots,n_m$. These lengths add up to n. Let c be the passing score for the overall test. The passing scores for the subtests are c_1, $c_2,\ldots,c_m$. As mentioned in earlier chapters, the subtest passing scores are set up such that their sum is the overall passing score.

Underlying the responses to the items is the latent trait which takes values in the sample space Ω. For the beta-binomial model, θ is the proportion of items in the pool that the subject answers correctly; thus θ ranges from 0 to 1. In latent trait models such as the Rasch, θ is the value of the unobservable latent variable which serves to explain the responses on the set of items. In general, θ is a unique function of the expected number of correct responses and ranges over the entire real line.

Let the overall test be administered to a population of subjects and let the probability density function (pdf) of the latent trait θ be $p(\theta)$. (For Bayesians, $p(\theta)$ represents the prior pdf associated with a given subject.) For the test score x, let $f(x)$ and $f(x|\theta)$ be the marginal pdf and the conditional pdf with respect to θ. Likewise the pdf's associated with each subtest score xj are denoted by $fj(x)$ and $fj(x|\theta)$. As in most latent trait models, the condition of local independence will be assumed to be satisfied; hence joint probabilities can be written as products of the relevant marginal probabilities.

It is now assumed that all subjects with scores on the overall test score smaller than c will be provided with remedial instruction. This additional instruction is given only on the objective(s) that the student has not mastered. In other words, remedial instruction will be given on the j-th objective if the subtest score xj is below the subtest passing score cj. With m as the number of objectives, the number of different remediation situations amount to 2^m-1. For example, there is one case where all the objectives are missed and m cases where the number of missed objectives is either 1 or $m-1$.

To form a complete solution for the budgetary problem posed in this chapter, a complete description of the cost of remediation would be required for each of the remediation situations. As an approximation to the reality of instruction, it is not unreasonable to assume that the cost remains essentially the same for each remediation situation involving a given number of objectives. This assumption requires that the objectives be about the same level of difficulty.

It will now be assumed that, for a subject with ability θ, the cost of remediation on k objectives can be described by a non-increasing function $df(\theta)$. Thus for the same number of objectives, remediation will cost more for less able students than it will for more able students.

For the subject with ability θ, let $Z_j(\theta)$ be the probability that the j-th objective has not been mastered. In other words,

$$Z_j(\theta) = \Pr(x_j < c_j|\theta). \tag{1}$$

Let $S_k(\theta)$ be the probability that the subject misses any k objectives among the m objectives. Then $S_k(\theta)$ is the symmetric sum

$$S_k(\theta) = \sum_{\underset{\sim}{u}} \prod_{j=1}^{m} \left(Z_j(\theta)\right)^{u_j}\left(1 - Z_j(\theta)\right)^{1-u_j}, \tag{2}$$

where the vector $\underset{\sim}{u} = (u_1,\ldots,u_m)$ of 0's and 1's extends over the region defined by $u_1 + \ldots + u_m = k$.

With $\delta_k(\theta)$ as the remediation cost associated with k objectives, the cost at the ability θ is expected to be

$$D(\theta) = \sum_{k=1}^{m} \delta_k(\theta)S_k(\theta). \tag{3}$$

Over a population of subjects where $p(\theta)$ is the pdf for θ, the expected cost at the overall passing score c is

$$\gamma(c) = \int_{\Omega} D(\theta)p(\theta)d\theta. \tag{4}$$

If the population consists of M subjects and if the passing score is selected as c (and hence the subtest passing scores are $c_1,\ldots,c_k$), the expected cost will be equal to $M\gamma(c)$.

3. Estimation of Parameters

By use of appropriate psychometric models, the functional forms for the probabilities $S_k(\theta)$ may be obtained. For example, if test scores follow the binomial model, then the pdf of x_j is

$$F_j(x_j|\theta) = \binom{n_j}{x_j}\theta^{x_j}(1-\theta)^{n_j-x_j}; \tag{5}$$

hence

$$Z_j(\theta) = \sum_{x_j<c_j} F_j(x_j|\theta). \tag{6}$$

By additionally assuming that the pdf $p(\theta)$ belong to some well-known family such as the beta family, this pdf can be approximated if there are enough subjects taking the test.

The specification of the various costs $\delta_k(\theta)$ probably would require careful deliberation and judgement. As a first approximation, the cost $\delta_k(\theta)$ may be taken to be proportional to the number k of objectives yet to be mastered. In addition, by imposing suitable functional forms on the $\delta_k(\theta)$, an approximation can be made which reflects the actual cost of remediation in real-life situations.

The next section provides an overall result for the case where the cost $\delta_k(\theta)$ is a linear function of k.

3. A General Result When $\delta_k(\theta)$ is Proportional to k

Consider the case where each cost function $\delta_k(\theta)$ is proportional to k, namely

$$\delta_k(\theta) = kh(\theta). \tag{7}$$

For this case, the expected cost at the ability θ (Equation (3)) is given as

$$D(\theta) = h(\theta)\left(S_1(\theta) + 2S_2(\theta) + \ldots + mS_m(\theta)\right).$$

We will show that $D(\theta)$ takes the following simple form

$$D(\theta) = h(\theta) \sum_{j=1}^{m} Z_j(\theta). \tag{8}$$

In fact, let the random variable B_j, $j=1,2,\ldots,m$ take the value 1 with probability $Z_j(\theta)$ and the value 0 with probability $1-Z_j(\theta)$. Then the sum $\sum_{j=1}^{m} B_j$ represents the number of objectives that the subject does not master. Since the expected value of each B_j is $Z_j(\theta)$, the expected value of the sum ΣB_j is the sum $\Sigma Z_j(\theta)$. It may be noted that the sum $S_1(\theta) + 2S_2(\theta) + \ldots + mS_m(\theta)$ is another form for the expected value of the sum ΣB_j.

It follows from the above remarks that as long as the cost is proportional to the number of objectives, computations for the expected costs due to remediation on multiple objectives will reduce to the simple case considered previously by Huynh (1980).

The following section illustrates the case of the beta-binomial with linear costs.

4. Special Case 1: The Beta-Binomial Model with Linear Costs

Consider now the beta-binomial model as defined by the following pdf's:

$$f(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}, \quad x=0,1,\ldots,n \tag{9}$$

and

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < \theta < 1. \tag{10}$$

The two parameters α and β may be estimated from sample data via one of several estimation techniques such as the moment procedure or the maximum likelihood procedure. Let $\bar{x}$ and s be the sample test score mean and standard deviation. In addition, let $\hat{\alpha}_{21}$ be the KR21 reliability coefficient as defined by

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right). \tag{11}$$

(In the case of a negative $\hat{\alpha}_{21}$, simply replace the value computed from Equation (11) by any positive reliability estimate.) The moment estimates for α and β are given as

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x} \tag{12}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n. \tag{13}$$

As in Section 2, let us presume that the overall n-item test is comprised of m subtests, each with $n_1,\ldots,n_m$ items. In addition, let the passing scores be $c_1,\ldots,c_m$ on these m subtests. Thus the probabilities $Z_j(\theta)$ defined in the previous section are given as

$$Z_j(\theta) = \sum_{x_j=0}^{c_j-1}\binom{n_j}{x_j}\theta^{x_j}(1-\theta)^{n_j-x_j}. \tag{14}$$

Moreover, let us consider the case where the costs of remediation take the forms

$$\delta_1(\theta) = h(\theta) = (\gamma_0-\gamma_1)(1-\theta) + \gamma_1$$

and

$$\delta_k(\theta) = kh(\theta).$$

It follows from Equation (8) that the expected remediation cost for a subject with true ability θ is

$$D(\theta) = \left[(\gamma_0-\gamma_1)(1-\theta) + \delta_1\right] \sum_{j=1}^{m} Z_j(\theta).$$

Over the population of subjects, the expected cost per student is given as

$$\gamma(c) = \int_0^1 D(\theta)p(\theta)d\theta$$

$$= \int_0^1 \left[(\gamma_0-\gamma_1)(1-\theta) + \delta_1\right]\left(\sum_{j=1}^{m} Z_j(\theta)\right) \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\, d\theta.$$

Thus

$$\gamma(c) = \frac{1}{B(\alpha,\beta)} \sum_{j=1}^{m} \sum_{x_j=0}^{c_j-1} \binom{n_j}{x_j} \left[(\gamma_0-\gamma_1)B(\alpha+x_j, n+\beta-x_j+1)\right.$$

$$\left. \gamma_1 B(\alpha+x_j, n+\beta-x_j)\right]. \qquad (15)$$

5. Special Case 2: The Rasch Model with Constant Costs

As indicated previously the Rasch model is used as the latent trait model for the analysis of the South Carolina BSAP data. In this model, the probability that a subject answers an item correctly is a function of the difference between his ability (θ) and the item difficulty (δ). This function takes th/ form

$$P(\theta) = \frac{e^{\theta-\delta}}{1 + e^{\theta-\delta}}.$$

The probability that corresponds to an incorrect response is therefore

$$Q(\theta) = \frac{1}{1 + e^{\theta-\delta}}.$$

Consider now an overall test of n items with item difficulties $\delta_1,\delta_2,\ldots,\delta_n$. Let the vector $A = (A_1,A_2,\ldots,A_n)$ denote the responses

to the n items. Each response a_j is either 0 or 1. For a subject with ability θ, the probability associated with the value $\underset{\sim}{a} = (a_1, a_2, \ldots, a_r)$ for the vector A is

$$P(A = \underset{\sim}{a}|\theta) = \prod_{j=1}^{n} \left(P_j(\theta)\right)^{a_j} \left(Q_j(\theta)\right)^{1-a_j} \tag{16}$$

where

$$P_j(\theta) = \frac{e^{\theta-\delta_j}}{1 + e^{\theta-\delta_j}}$$

and

$$Q_j(\theta) = \frac{1}{1 + e^{\theta-\delta_j}}.$$

It may be noted that Equation (16) can be written as

$$P(A = \underset{\sim}{a}|\theta) = \left(\prod_{j=1}^{n} Q_j(\theta)\right) \prod_{j=1}^{n} \left(\frac{P_j(\theta)}{Q_j(\theta)}\right)^{a_j}.$$

Thus, by letting

$$H = \prod_{j=1}^{n} Q_j(\theta)$$

and

$$\xi_j(\theta) = \frac{P_j(\theta)}{Q_j(\theta)}, \quad j = 1,2,\ldots,n,$$

it may be noted that

$$P(A = \underset{\sim}{a}|\theta) = H \prod_{j=1}^{n} \left(\xi_j(\theta)\right)^{a_j}. \tag{17}$$

When the test items have been calibrated (e.g., when all the item difficulty parameters δ_j are known), the pdf associated with the raw score $x = \Sigma a_j$ at each ability θ is given as

$$f(x|\theta) = \sum_{\Sigma a_j = x} P(A = \underset{\sim}{a}|\theta)$$

or

$$f(x|\theta) = H \sum_{\Sigma a_j = x} \prod_{j=1}^{n} \left(\xi_j(\theta)\right)^{a_j}. \tag{18}$$

In order to compute the probability $f(x|\theta)$ of Equation (18), let us follow the notation used by Gustafsson (1980) and denote

$$\gamma_x(\xi_1,\xi_2,\ldots,\xi_n) = \sum_{\Sigma a_j = x} \prod_{j=1}^{n} \left(\xi_j(\theta)\right)^{a_j}, \tag{19}$$

so that

$$f(x|\theta) = H\gamma_x(\xi_1,\xi_2,\ldots,\xi_n).$$

The γ_x functions of Equation (19) may now be computed via the following recursive formula reported in Fischer (1974, p. 250)

$$\gamma_x(\xi_1,\ldots,\xi_t) = \gamma_x(\xi_1,\ldots,\xi_{t-1}) + \xi_t\gamma_{x-1}(\xi_1,\ldots,\xi_{t-1}) \tag{20}$$

where $0 \leq x \leq t$ and $t = 1,\ldots,n$.

As pointed out in Gustafsson (1980, p. 381), this formula can be applied recursively to compute the probability associated with each raw score x. Starting with $\gamma_1(\xi_1) = \xi_1$ and $\gamma_0(\xi_1) = 1$, one more variable can be added so that

$$\gamma_1(\xi_1,\xi_2) = \gamma_1(\xi_1) = \xi_2\gamma_0(\xi_1) = \xi_1 + \xi_2$$

and

$$\gamma_2(\xi_1,\xi_2) = \gamma_2(\xi_1) + \xi_2\gamma_1(\xi_1) = 0 + \xi_2\xi_1 = \xi_1\xi_2.$$

Likewise, with one additional variable, we have

$$\gamma_1(\xi_1,\xi_2,\xi_3) = \gamma_1(\xi_1,\xi_2) + \xi_3\gamma_0(\xi_1,\xi_2) = \xi_1 + \xi_2 + \xi_3,$$

$$\gamma_2(\xi_1,\xi_2,\xi_3) = \gamma_2(\xi_1,\xi_2) + \xi_3\gamma_1(\xi_1,\xi_2)$$

$$= \xi_1\xi_2 + \xi_3(\xi_1 + \xi_2) = \xi_1\xi_2 + \xi_1\xi_3 + \xi_2\xi_3,$$

and

$$\gamma_3(\xi_1,\xi_2,\xi_3) = \gamma_3(\xi_1,\xi_2) + \xi_3\gamma_2(\xi_1,\xi_2) = 0 + \xi_3\xi_1\xi_2 = \xi_1\xi_2\xi_3.$$

The computation scheme described by the recursive formula (20) may be used to compute the conditional probability $f_j(x_j|\theta)$ associated with the j-th subtest. The probabilities $Z_j(\theta)$ and $S_k(\theta)$ of Section 2 can then be computed, and with the specification of the cost functions $d_k(\theta)$ and the density $p(\theta)$, the expected cost per student $\gamma(c)$ can also be computed for each passing score c.

As pointed out in previous chapters, for a test with n previously calibrated items, there are n+1 separate values for the true ability latent trait. Each value corresponds to a given raw score. Strictly speaking, the zero raw score corresponds to the true ability $\theta_0 = -\infty$ and the perfect score raw score n corresponds to the true ability $\theta_n = +\infty$. However, to facilitate various computations, both θ_0 and θ_n have been equated to two finite values obtained by suitable linear extrapolation.

If historical data exist which provide the relative frequency $p(\theta)$ at each ability θ, then the marginal probabilities associated with missing 1,2,...,m objectives can be computed. More specifically, the probability of missing k objectives is given by the sum

$$W_k = \sum_{h=0}^{n} S_k(\theta_h)p(\theta_h).$$

If costs are constant across students, then the expected cost for each subject is equal to the sum

$$\sum_{k=1}^{m} d_k W_k.$$

6. An Illustration for the Rasch Model with Constant Costs

To illustrate the use of the Rasch model in studies of budgetary implications, let us consider the BSAP math test for grade two administered in 1981. The test was calibrated on the basis of approximately 2600 students and the item difficulty estimates (listed in Table 7) are reproduced in Table 34 of this chapter. As previously mentioned in Chapter 1, the passing score for the overall test was set at 22. The test consists of five objectives, namely OP, CN, GE, ME, and PS, and their cutoff scores were set as 4, 4, 5, 5, and 4 via the Rasch constant-sum procedure.

With a total of 30 items, the overall test score ranges from 0 to 30. The Rasch ability at each raw score may be computed via the numerical approximation described in Chapter 1. The left side of

TABLE 34

Rasch Item Difficulty Parameters
for Grade 3 Math Test in 1981

Item	Difficulty	Item	Difficulty	Item	Difficulty
OP02	-0.115	CN05	1.064	ME04	-0.006
OP07	0.765	CN20	-0.507	ME15	-1.791
OP08	0.497	GE01	0.041	ME20	0.501
OP12	-0.016	GE08	-0.138	ME21	1.793
OP15	0.612	GE09	-1.265	PS06	-0.473
OP20	1.179	GE18	-1.516	PS11	-0.264
CN16	-0.177	GE19	-1.094	PS12	0.388
CN13	0.821	GE20	0.415	PS13	0.175
CN08	-0.658	ME01	-1.003	PS17	0.471
CN01	0.377	ME08	-0.766	PS21	0.688

Table 35 reports the Rasch ability at each raw score along with the number of students having this raw score.

The right side of Table 35 reports the probabilities $S_k(\theta)$ that a student with each Rasch ability value θ will not master any of the $k = 1, 2, 3, 4,$ or 5 math objectives. The last line of Table 35 reports the mean W_k of each probability $S_k(\theta)$ weighted according to the number of students.

A variety of computations for remediation costs may be performed from the probabilities in Table 35. For example, if the remediation costs are constant across students and equal to d_k for k objectives, then the projected cost of setting the overall passing score at 22 is the sum $\sum_{k=1}^{5} d_k W_k$. As an illustration, letting $d_1 = 10$, $d_2 = 15$, $d_3 = 18$, $d_4 = 20$, and $d_5 = 21$, then the projected remediation cost per student is 9.51.

7. Allocation of Resources to Schools

In a number of situations, resources are available at the state level which need to be allocated to each school district within the state for the purpose of instructional remediation. If instructional remediation is to be carried out at the objective level and if the cost remains constant across students and objectives, then the

TABLE 35

List of Probabilities $S_k(\theta)$ of Missing k Objectives

Raw Score	Rasch Ability	Number of Students	$S_k(\theta)$ at k of 1	2	3	4	5
0	-4.747	1	0.000	0.000	0.000	0.000	1.000
1	-3.684	0	0.000	0.000	0.000	0.000	1.000
2	-2.928	1	0.000	0.000	0.000	0.000	1.000
3	-2.459	1	0.000	0.000	0.000	0.001	0.999
4	-2.107	3	0.000	0.000	0.000	0.004	0.996
5	-1.820	5	0.000	0.000	0.000	0.011	0.989
6	-1.572	9	0.000	0.000	0.000	0.024	0.976
7	-1.351	17	0.000	0.000	0.001	0.046	0.953
8	-1.150	58	0.000	0.000	0.003	0.080	0.917
9	-0.963	97	0.000	0.000	0.007	0.127	0.866
10	-0.787	202	0.000	0.001	0.017	0.186	0.797
11	-0.619	329	0.000	0.002	0.035	0.253	0.711
12	-0.457	523	0.000	0.006	0.064	0.319	0.620
13	-0.299	780	0.001	0.015	0.108	0.375	0.501
14	-0.144	1044	0.003	0.032	0.166	0.409	0.390
15	0.010	1295	0.008	0.062	0.231	0.413	0.286
16	0.163	1578	0.019	0.107	0.293	0.384	0.196
17	0.317	1844	0.040	0.168	0.338	0.328	0.123
18	0.474	2249	0.076	0.237	0.353	0.254	0.071
19	0.634	2455	0.131	0.302	0.333	0.177	0.036
20	0.799	2681	0.203	0.345	0.280	0.110	0.017
21	0.973	2990	0.286	0.354	0.208	0.059	0.006
22	1.156	3135	0.362	0.321	0.135	0.027	0.002
23	1.353	3384	0.413	0.255	0.074	0.010	0.001
24	1.569	3517	0.420	0.173	0.033	0.003	0.000
25	1.812	3677	0.375	0.097	0.012	0.001	0.000
26	2.095	3760	0.288	0.043	0.003	0.000	0.000
27	2.441	3728	0.181	0.013	0.000	0.000	0.000
28	2.904	3475	0.085	0.002	0.000	0.000	0.000
29	3.654	2707	0.021	0.000	0.000	0.000	0.000
30	4.711	1435	0.003	0.000	0.000	0.000	0.000
		Weighted Mean =	0.202	0.147	0.118	0.093	0.062

allocation of funds can be carried out on the basis of the total number of nonmastered objectives by all students in the district.

On the other hand, if the remediation cost varies according to the ability level of the student and the complexity of the objective, then the cost functions $d_k(\theta)$ may be specified at the state level and the average cost per student may then be computed for each school district. The allocation of budgeted remediation funds to each school district may then be made proportional to the number of students and the local cost per school.

References

Fischer, G. H. Einführung in die Theorie psychologischer Tests. Grundlagen und Anwendungen. Bern: Huber, 1974.

Gustafsson, J. E. A solution of the conditional estimation problem for long tests in the Rasch model for dichonomous items. Educational and Psychological Measurement, 1980, 40, 377-385.

Huynh, H. Budgetary consideration in setting passing scores. In Huynh, H., and Saunders, J. C., Solutions for some technical problems in domain-referenced mastery testing. Final Report of Grant NIE-G-78-0087, August, 1980.

CHAPTER 8

ASSESSING THE BUDGETARY IMPACT OF REMEDIATION IN BASIC SKILLS ASSESSMENT PROGRAMS: II. INSTRUCTIONAL CONSIDERATION

1. Introduction

In the construction of multiple-choice test items for a basic skills assessment program, considerable emphasis is put on the selection of distractors which reflect major types of errors. When this is done, the responses to the test items reveal not only the overall performance of the student but also the major types of errors which may need remedial instruction.

The seriousness of each error is probably a direct function of the amount of remedial instruction needed to correct it. Some errors are easy to overcome; others may demand more effort. Thus, in the allocation of funds to schools or school districts for remedial instruction, perhaps one needs to consider not only the total number of students who do not meet the passing score and the number of non-mastered objectives, but also the seriousness of the errors made by these students.

This chapter provides an illustration of how the level of complexity in remediation can be taken into account in the process of budget allocation.

2. An Index for the Seriousness of Errors on a Test

Consider now a test which is comprised of n multiple-choice items. For the i-th item, let k_i be the number of alternatives; k_i may vary from item to item. For a group of students who do not meet the minimum level of performance, let m_{ij} be the number of students who choose the j-th option on the i-th item.

Let us assume that it is possible to quantify the seriousness of all the errors displayed in the incorrect options of the multiple-choice items on a *common scale*. This common scale extends from zero to a convenient maximum value C. Since the correct options of the

multiple-choice items do not involve any error, their level of seriousness may be equated to zero on this scale.

For the i-th item, let c_{ij}, $j=1,\ldots,k_i$ be the seriousness level of the j-th option. With M as the total number of incorrect responses to the items of the test, the seriousness of error for the entire test may be taken as

$$\varepsilon = \left(\sum_{i=1}^{n} \sum_{j=1}^{k_i} n_{ij} c_{ij} \right) / (MC) .$$

This index varies from 0 to 1. For a given group of students, ε approaches 0 when all the individual levels of seriousness are close to zero. On the other hand, when all these levels are near the maximum value C, ε will approach 1.

3. First Illustration: Comparing the Seriousness Level of the Reading Objectives of Grade Six

As mentioned in several previous chapters, the South Carolina Basic Skills Assessment Program (BSAP) consists, in part, of the administration of the basic skills tests in reading and math to several grade levels. For the reading test, the six objectives are decoding and word meaning (DW), main idea (MI), details (DE), analysis of literature (AL), reference usage (RE), and inference (IN). Each objective is measured by a six-item subtest; hence the reading test has 36 items altogether. For the sixth grade reading test administered in 1981, the passing score was set at 22.

As an illustration of the use of the index ε, let us focus on the sample of students used in the setting of the passing score in the 1981 test administration. In the sample, there are 938 students who score below the passing score of 22. The left part of Table 36 reports the number of students in each option of the 36 multiple-choice items. (For each item, there are a small number of students with no response or unrecognized responses; these are not listed in Table 36.)

The right part of Table 36 reports the seriousness level assigned to each incorrect option on a scale from 0 to 5. The rating was done

TABLE 36

Data for the Illustration of Section 3

Item Number	Frequency in Option				Seriousness of Option			
	1	2	3	4	1	2	3	4
1	343	76	68	451	3	1	1	0
2	188	661	60	26	2	0	3	4
3	34	47	70	778	5	4	3	0
4	292	350	112	174	1	0	1	4
5	95	667	62	107	3	0	5	4
6	366	118	87	352	5	3	4	0
7	177	112	239	408	0	4	5	3
8	296	312	198	119	3	0	5	4
9	345	135	202	251	0	5	3	4
10	201	279	338	119	5	3	0	4
11	334	194	253	155	0	5	4	3
12	317	257	162	188	0	5	4	4
13	456	203	183	96	0	1	1	1
14	79	211	560	82	3	3	0	3
15	107	232	131	459	2	2	2	0
16	184	519	137	86	2	0	2	2
17	166	423	160	183	3	0	3	3
18	235	226	76	391	3	3	3	0
19	191	313	128	305	2	2	2	0
20	253	224	232	217	2	0	2	2
21	206	192	283	236	4	4	0	5
22	338	276	171	141	5	0	4	4
23	241	205	217	263	0	3	3	3
24	306	175	242	203	3	3	3	0
25	128	172	509	126	3	3	0	3
26	79	103	631	114	3	3	0	3
27	175	63	112	585	4	3	3	0
28	78	100	703	54	4	3	0	4
29	41	178	678	31	4	3	0	4
30	134	125	471	188	5	4	0	4
31	319	310	138	165	0	3	4	4
32	134	249	270	267	4	3	0	3
33	225	393	186	111	4	0	5	4
34	213	258	230	215	0	3	3	5
35	194	270	254	203	3	3	0	3
36	225	279	165	244	4	4	0	4

by two school teachers; a diary of their discussion about the seriousness of each error is included in the appendix to this chapter. (Test security does not permit detailed descriptions of the test items.)

Using the data in Table 36, the index for the seriousness of errors was found to be .607 for DW, .723 for MI, .643 for DE, .637 for AL, .647 for RE, and .662 for IN. Thus, for the situation under consideration, the error seriousness for each objective may be listed from the least serious to the most serious as DW, AL, DE, RE, IN, and MI.

4. Second Illustration: A Consideration for Equitable Budget Allocation for Remedial Instruction

The assessment of the seriousness level of the subtests for the objectives as illustrated in the previous section may help to equitably allocate the budget for instructional remediation to schools or school districts. When instructional remediation is to be given to each non-mastered objective, the formula for budget allocation perhaps should be based on the total number of cases in which each objective is missed and the level of seriousness of this objective.

For example, let us consider the allocation of remediation funds to k schools. The funds are to be used for sixth graders who do not meet the passing score of 22. Let us assume also that remediation is conducted for each of the objectives missed by a student. Let m_{ij}, $i = 1,\ldots,k$ and $j = 1,2,\ldots,6$ be the number of students in the i-th school who missed j objectives. In addition, let ε_j, $j = 1,2,\ldots,6$ be the level of seriousness of the errors associated with the j-th objective. Then the impact of remediation on the i-th school may be taken as the sum

$$I_i = \sum_{j=1}^{6} m_{ij}\varepsilon_j .$$

An equitable budget allocation may then be accomplished by dividing the total funds to each school in proportion to the indices I_i.

Numerical Example

Consider the allocation of remediation funds to four schools. For each school, the number of times that each objective is missed by a student is listed in Table 37. Granting that the seriousness of each objective is given as in Section 3, the indices I_i are 89.5, 153.5, 107.8, and 158.0 for the schools A, B, C, and D respectively. If a sum of \$100,000 is available, then, via the I_i index, each of these schools would be allocated \$17,590, \$30,169, \$21,187, and \$31,054 respectively. Had the budget consideration been on the basis of the total number of missed objectives (last column of Table 37), the funds allocated to the schools would have been \$17,413, \$30,346, \$21,639, and \$30,602 respectively.

TABLE 37

Number of Times Each Objective Is Missed

	Missed Objective						
School	DW	MI	DE	AL	RE	IN	Total
A	20	30	15	28	16	27	136
B	12	17	40	87	59	22	237
C	60	14	22	29	30	14	169
D	18	57	82	22	43	17	239

5. Use of ε to Assess Instructional Equivalence of Test Forms

Though the index ε for the seriousness of errors is proposed for studies of the impact of remediation in budget consideration, it may also be used to assess the instructional equivalence of various forms of a given test. When testing is carried out for diagnostic purposes, content validity and the seriousness of the errors portrayed in the multiple-choice items are perhaps of major importance. If this is the case and if alternate forms are needed, these forms must display the same content area as well as the same level of seriousness in the errors which are to be remediated. By using the ε index, one may assert whether these alternate forms are equivalent in terms of the complexity of the errors which need further instruction.

Appendix

Summary of the Teachers' Discussion Regarding the Seriousness of the Options in the Grade Six Reading Test

Decoding and Word Meaning (DW)

1.
3	A.	sound similarity
1	B.	random
1	C.	random
0	D.	correct response

In option A the child is confusing wealthy with healthy. Remediation would require review of initial consonant sounds. Options B and C could be remediated by stressing proper testing procedures and discouraging guesswork.

2.
2	A.	response to context
0	B.	correct response
3	C.	response to context
4	D.	response to context

The difficulty in remediating options A, C, and D is directly related to the plausibility of the answer.

3.
5	A.	opposite
4	B.	response to context
3	C.	response to context
0	D.	correct response

Option C is a plausible response which might seem logical to a child. It can be remediated by stressing the need to read for details. Options A and B are both possible results of ingrained substandard speech patterns. However, option A, as a direct opposite, would be more difficult to clarify.

4.
1	A.	response to context
0	B.	correct response
1	C.	response to context
4	D.	structural similarity

Options A and C would seem plausible <u>if</u> the child did not read the <u>entire</u> selection. This could be remediated by emphasis on careful reading. Option D is a random choice based on similar structure of the two words without careful reading. The child must be taught that word similarity need not be related to meaning.

5. 3 A. response to base without regard to affix
   0 B. correct response
   5 C. opposite
   4 D. response to affix

In option A the child responded to the base word. The child who so responds has mastered the concept of word-base. Remediation requires a review of affix meanings and usage. In option D the child responded to the affix rather than the base. This shows he has not yet mastered the concept of word-base. Remediation requires a review of the nature of word structure including both base and affix. Option C may be the result of confusion of affix meanings. On the other hand, it may result from total lack of knowledge of the word. Reason for the mistake must be determined before remediation can begin.

6. 5 A. opposite
   3 B. response to base without regard to affix
   4 C. random choice
   0 D. correct response

In option B the child responded to the base word. The child who so responds has mastered the concept of word-base. Remediation requires a review of affix meaning and usage. In option C the child did not know the word-base or meaning. Remediation requires vocabulary building. Guesswork should be discouraged. The opposite meaning in option A is a result of disregarding the affix. Remediation requires a more extensive review of affix meaning and word structure.

Main Ideas (MI)

7. 0 A. correct response
   4 B. unsupported
   5 C. contradicted
   3 D. narrow scope

8. 3 A. narrow scope
   0 B. correct response
   4 C. unsupported
   5 D. contradicted

9. 0 A. correct response
   5 B. contradicted
   3 C. narrow score
   4 D. unsupported

10. 5 A. contradicted
    3 B. narrow scope
    0 C. correct response
    4 D. unsupported

11.	0	A.	correct response
	5	B.	contradicted
	4	C.	unsupported
	3	D.	narrow scope

12.	0	A.	correct response
	5	B.	contradicted
	4	C.	unsupported
	4	D.	unsupported

A statement of narrow scope focuses on a minor detail of the selection rather than the main idea. Remediation would include reviewing the concepts of main idea and supporting ideas, possibly by use of outlining the selection.

The selections do not contain sufficient evidence to corroborate unsupported statements. Remediation would emphasize reading material for accuracy.

Contradictive statements express ideas opposite to those in the selections. This exemplifies minimal reading comprehension. Remediation would require extensive "reading for meaning" exercises.

Details (DE)

13.	0	A.	correct detail
	1	B.	incorrect detail
	1	C.	incorrect detail
	1	D.	incorrect detail

14.	3	A.	incorrect detail
	3	B.	incorrect detail
	0	C.	correct detail
	3	D.	incorrect detail

15.	2	A.	incorrect detail
	2	B.	incorrect detail
	2	C.	incorrect detail
	0	D.	correct detail

16.	2	A.	incorrect detail
	0	B.	correct detail
	2	C.	incorrect detail
	2	D.	incorrect detail

17.	3	A.	incorrect detail
	0	B.	correct detail
	3	C.	incorrect detail
	3	D.	incorrect detail

18. 3 A. incorrect detail
3 B. incorrect detail
3 C. incorrect detail
0 D. correct detail

In each selection, all the given options are mentioned. There is a key word or phrase in each stimulus to which the child should respond. Since a major cause of failure to recognize important details is reading too fast, remediation should include teaching the child to read for comprehension rather than speed. "Reading for meaning" activities and vocabulary development should be included in remediation. In selections 14, 17, and 18 remediation should also include training and exercises in the use of sequence skills.

Analysis of Literature (AL)

19. 2 A. opinion
2 B. opinion
2 C. opinion
0 D. fact

20. 2 A. fact
0 B. opinion
2 C. fact
2 D. fact

There is no varying degree of difficulty of remediation for the incorrect responses to items 19 and 20. The problem involved is the inability to distinguish between fact and opinion. The child needs to be taught the difference between subjective and objective reasoning. Mastery of these reasoning skills would require extensive reading practice using selections similar to these test items.

21. 4 A. inaccurate description of plot
4 B. inaccurate description of plot
0 C. accurate description of plot
5 D. accurate description of character

In option A the child has made the mistake of focusing on a detail in the selection. Option B is unrelated to the selection. Although the child may have an understanding of plot, his lack of comprehension skills caused him to select an inaccurate description of plot. Remediation would encompass exercises in reading comprehension.

If a child has not mastered the concept of plot, he may choose option D. This presents a more serious remediation problem involving the basic elements of literary composition.

22. 5 A. accurate description of plot
0 B. accurate character description
4 C. inaccurate character description
4 D. inaccurate character description

Options C and D are inaccurate descriptions of Elizabeth's character. Remediation for both of these options would involve more careful reading with attention to detail.

Option A requires more extensive remediation because it shows non-mastery of the concept of character, which is a basic element of literary composition.

23. 0 A. onomatopoeia
3 B. no
3 C. no
3 D. no

24. 3 A. no
3 B. no
3 C. no
0 D. simile

Onomatopoeia and simile are figures of speech. An incorrect response on item 23 or 24 would indicate that the child is not yet able to recognize the stated figure of speech in context. Remediation in both cases involves further exposure to these figures of speech.

Reference Usage (RE)

25. 3 A. incorrect reference source
3 B. incorrect reference source
0 C. correct reference source
3 D. incorrect reference source

Options A, B, and D are incorrect reference sources and show a lack of understanding of what a call number is. Remediation would involve the teaching of library organization and the selection of reference sources.

26. 3 A. incorrect reference source
3 B. incorrect reference source
0 C. correct reference source
3 D. incorrect reference source

In selecting a reference source for this item, the child should note the key word "pictures" in the stimulus. Remediation of options A, B, and D would involve stressing the importance of reading for

information with attention to details. Additional remediation on the use of reference books such as encyclopedias and dictionaries would be helpful.

27. 4 A. incorrect response
5 B. incorrect response
3 C. incorrect response
0 D. correct response

Choice of option C shows some thought. Remediation would involve a discussion of topics and subtopics. Choice of option A shows possible inattention to detail. Remediation would involve reading carefully with attention to detail. Choice of option B would indicate a total lack of understanding of the use of a table of contents. Remediation would entail a complete review of the use of a reference source.

28. 4 A. incorrect response
3 B. incorrect response
0 C. correct response
4 D. incorrect response

The child who selected option B probably understood how to use a chart. His mistake was likely a result of inattention to detail. Remediation would employ practice in the reading and use of charts.

Remediation of options A and D would be more difficult since a child who selected one of these options may lack a basic understanding of how to read and use a chart.

** It may be that the use of chemical symbols could confuse some children who were actually familiar with chart usage.

29. 4 A. incorrect response
3 B. incorrect response
0 C. correct response
4 D. incorrect response

A child who selected option B would have a fair understanding of the use of a card catalog, but did not read all the options carefully. Remediation would entail extensive practice in the use of the card catalog.

Remediation for options A and D would require a thorough review of the card catalog. Some remediation in spelling might be useful.

30. 5 A. incorrect response
4 B. incorrect response
0 C. correct response
4 D. incorrect response

In choosing option D the child responded to the stimulus by finding the correct topic but neglected to find the correct subtopic. In responding with option B, the child found the correct subtopic under the incorrect topic. While the reason for each mistake was different, they would be equally difficult to remediate. Remediation would involve review of topic and subtopic. In choosing option A, the child showed non-mastery of topic and subtopic. Remediation would follow the same lines as that for D and B but would be more extensive.

** The index sample used in item 30 may possibly have contributed to the confusion of those children who made wrong choices.

Inference (IN)

31. 0 A. correct comparison
3 B. incorrect comparison
4 C. incorrect comparison
4 D. incorrect comparison

Option B is the easiest mistake to remediate. The child should be encouraged to read all descriptive materials carefully before making a decision. Option D shows inattention to detail since the child has chosen an answer which directly contradicts the stimulus. Option C shows a misunderstanding of the materials due also to inattention to details. Remediation would require much more reading practice with emphasis on attention to detail.

32. 4 A. contradicted cause
3 B. less likely cause
0 C. most likely cause (correct)
3 D. less likely cause

While options B and D are true statements, they do not respond directly to the stimulus. Remediation would entail practice in logical thinking with emphasis on the relationship between cause and effect. On the other hand, option A makes a totally untrue statement. Remediation would include exercises in reading with attention to detail and discussion of the material read as well as a review of the relationship between cause and effect.

33. 4 A. contradictory
0 B. most likely cause
5 C. unrelated statement
4 D. contradictory

Remediation of options A and D would involve reading practice with attention to key phrases. Option C would require much more extensive remediation along the same lines. There is little in the article to support such a conclusion. Therefore a child who chose this option would also need more instruction and exercises in logical reasoning.

34. 0 A. most reasonable conclusion
3 B. unsupported conclusion
3 C. unsupported conclusion
5 D. contradicted conclusion

There is lack of information to support options B and C. Remediation would involve teaching a child to draw conclusions based on adequate information. The fact that in option D the child has chosen a contradictory statement shows that he/she requires extensive reading practice with attention to drawing conclusions based on fact.

35. 3 A. less reasonable
3 B. less reasonable
0 C. most reasonable conclusion
3 D. less reasonable

Options A, B, and D are equally difficult to remediate because in each case the child drew a conclusion which was unsubstantiated by the selection. Remediation would involve reading practice with attention to drawing conclusions based on fact.

36. 4 A. incorrect outcome
0 B. correct outcome
4 C. incorrect outcome
4 D. incorrect outcome

Since options A, C, and D are clearly unreasonable outcomes, the child needs remediation in reading, deductive reasoning, drawing conclusions, and predicting outcomes. Exercises might include group discussion and asking open-ended questions.

PART E

SOME VIEWS ON PSYCHOMETRIC ISSUES

CHAPTER 9

A VIEW ON FUTURE PSYCHOMETRIC ISSUES IN MENTAL MEASUREMENT

1. How Far Have We Come in Classical Measurement Theory?

We have indeed come a long way since the dawn of this century, when someone cared enough to put down the equation which says that an observed test score has two additive parts, one reflecting the true ability of the examinee and the other summarizing various random factors conveniently referred to as error of measurement. We then make several statistical assumptions about the nature of this type of error and its relationship with the examinee's true ability. These assumptions have rendered us ample opportunity to study basic concepts such as reliability, standard error of measurement, validity, and the like and to learn of the appropriate ways to estimate them. Of course, the very basic assumptions in the classical approach to mental measurement presume that testing is done to a group of examinees; hence, the interpretation of test results is to be accomplished within the framework of a given group of examinees. In addition, concepts which directly affect the selection of test items such as item difficulty and item discrimination have to be defined for a particular population of examinees for which the test is intended.

The population-dependent characteristics of items, tests, and test score interpretation have come to bother many of us a great deal. If this is not your case, of course it was Fred Lord's, whose towering reign over mental measurement has been and will be felt for many years to come. I still remember the cold days at Iowa and the agony of referring to p-values as item difficulty and point-biserials as item discrimination and accepting the fact that these item characteristics vary from population to population.

Transcript of a talk given by Huynh Huynh as part of the symposium "Future Directions for Mental Measurement." New York: Meetings of AERA and NCME, March 19, 1982.

I have had and probably will have students come to ask me about estimating test reliability in a pretest-posttest design. Always I have told them not to combine the pretest and posttest data in reliability estimation because the concentration of pretest scores at one place and of posttest data at a second place would result in an estimate which is very close to one. "But then what should I do?" I should confess that I have not come up with any satisfactory answer.

2. Have We Beaten the Linear Model for Test Scores to Death?

With the Hoyt discovery that the Kuder-Richardson Formula 20 reliability can be deduced from a two-way analysis of variance, there have been numerous studies using a multitude of linear decompositions for test scores. These studies are, of course, exciting because they provide ways to identify various sources of errors which have been lumped in one pot called error of measurement. Dependability and generalizability are the name of the game.

While I have always admired the beauty of the analysis of variance (and have messed around with it in the context of repeated measures for a while) and have no doubt of its usefulness in describing the behavior of test data for a particular population, I feel somewhat uneasy seeing its forces pushed on the modeling of item responses. When item responses are simply coded as zero or one, I wonder how we can explain with eyes open that zero is actually the sum of a number of uncorrelated components and one is also constituted of a number of unrelated parts.

3. So We Want to Look at the Item by Itself: Why Not Insist on the Simple Rasch Model?

If we are not interested in item parameters which are population-dependent, of course we have to look for item parameters which are population-independent. Here come latent trait and item response models. The beauty of these models has been recognized for some time, but their full force did not venture into the educational

testing enterprise until recently, when fast computers allowed the execution of complex estimation processes.

We have at hand a variety of latent trait models from which to choose. There are those of us who categorically argue for the Rasch model; the reasons are that they provide "objective measurements" like those in the physical world (length, time, and mass) and that the estimation of item parameters can be accomplished independently of the examinee's test score. There are others who accept the more complex three-parameter logistic model because it provides more flexibility in explaining the observed item responses.

But then, do we really have "objective measurements" in the physical world? Is there a device which is called a universal ruler that will give us an objective measure for the length of a table? Perhaps yes, perhaps no. We are in constant motion in space and time, and those of us who appreciate the beauty of the pioneering work of Albert Einstein are probably still fascinated by the interaction between time and space, a phenomenon researchers in the subatomic world have not ignored. Perhaps some day we will be able to map the complexity of the human mind into a finite number of dimensions; as for myself, I still believe in the infinity of that white substance, in terms of both its rationales and contradictions, and have never been sure about how that world of boundlessness would find accommodation with as simple and finite a thing as a test item.

Most of us have been in contact with estimation concepts such as unbiasedness, sufficiency, and maximum likelihood and have learned to use them carefully. Although these concepts are inventions of the very best of the statistical mind, they do not always provide answers which are intuitively justified; so the insistence on a particular model because it has some desirable estimation property may not be the best course of human judgement. I still remember the first time the normal distribution was introduced with all its simplicity and ease in estimation. Here the population mean can be estimated by the sample mean; here the population variance can be

estimated by the sample variance, and the two estimates are independent of each other. Upon further study, I came to realize that the normal distribution is the only situation in which this type of independence exists. Of course, there are so many data sets which do not conform to the famous curve discovered by Laplace and Gauss many years ago, and of course we cannot ignore them because we have not yet achieved independent estimates for the unrelated traits in the data.

4. Is There Life Beyond the Three-Parameter Logistic Model?

Perhaps I have conveyed the feeling that I do not like the Rasch model. Oh, no. The Rasch model is indeed a simple and, in many ways, a powerful model which has done many of us great service. But in sorting through the many technical problems concerning the South Carolina Basic Skills Assessment Program, I realized that concerns for content had to take priority over statistical goodness of fit. So I have had to lay low my zeal for internal consistency so that we can move on with all the data cranking. Of course, we can justify the use of any model by carefully indicating that what we are doing is only approximating and that some day, when more knowledge has accumulated, we will do better in mapping the many intellectual traits that are dear to us.

Perhaps the intellectual world is more complex than the three-parameter logistic model. Why not attempt to think multivariate in latent trait theory? Oh, yes, there is research in this area now. Why not get away from using a parametric framework for the description of the item responses? Perhaps we need to make only the smallest number of assumptions and let the best of you in the audience take care of all the details.

Why not weaken the assumption of independence in the item responses to a variation of exchangeability? Why not simply require that various item characteristic functions be monotonic? A great

deal of research has been done in monotone regression; these results have yet to be expanded to the area of psychometrics. There is always a longing to have estimates for item parameters which are in some sense unrelated to estimates for the examinees. Why not try the switched-sample procedure in which the sample is divided into, say, two smaller samples, one subsample to be used primarily to estimate item parameters and the other subsample for ability estimation? Since item estimation and ability estimation are accomplished through two independent samples, perhaps all kinds of nice properties like consistency can be documented.

5. Can We Achieve Objective Measurement Through Order Constraints?

Perhaps there may be a time that we would feel comfortable to insist that a given test must consist of items for which the item characteristic curves do not cross. There is no reason to insist on the Rasch model; many models can do this as long as we impose some restrictions on the item parameters. Estimation of the item parameters would then be approached from something like monotone regression or maximum likelihood under order constraints. This may be a difficult problem, but a great number of details have been worked out in mathematical programming. Perhaps now is the time when we take a second look at various forms of item characteristic curves, impose suitable order restrictions, and adjust canned computer programs like LOGIST to these constraints.

6. Why Not Reformulate Generalizability Theory Within the Context of Item Response Theory?

Now, if you are unhappy with linear decomposition with 0-1 item data, why not try to formulate a generalizability theory within a context of item response or multiple contingency table? Statisticians have linearly decomposed the log of the likelihood function for years; why not we in mental measurement theory? Values of the log likelihood vary from minus infinity to zero, so at least we can feel at ease in cutting the log likelihood in small pieces.

7. On the Interpretation of Test Scores: Is There a Test Out There by the Name "Criterion-Referenced"?

The educational measurement enterprise has been deluged with endless writings on criterion-referenced testing in recent years. (The pace has slowed down a bit, though.) I fully understand the need to gear testing to very well defined educational objectives and the need to interpret test scores within the context of individual achievement, but is there a test which really bears the meaning of the term "criterion-referenced" as originally proposed by Glaser a few years ago?

Perhaps not. The Glaser definition for a criterion-referenced test attaches an absolute interpretation to test scores; so by looking at the test score, we can infer what a student can do or cannot do. I spent some time in Pittsburgh in the summer of 1973 pondering the definition and trying to see how a psychometric theory could be formulated for this type of test. If I took the Glaser definition seriously, then I would have to accept the assumption that student performance constituted a linear ordering (or hierarchy). This is consistent with the linear order implicit in the system of real numbers. However, how many educational accomplishments can actually be put in a linear sequence?

8. On the Interpretation of Test Scores: What Can We Accomplish via Decision Theory?

Most measures of educational performance are collected because some decisions need to be made. This is particularly true in testing programs which are designed for instructional purposes. So how do we formulate a psychometric theory which takes into account this type of conceptualization?

Two questions may be raised. First, what are some of the best ways to tap the information contained in the test data? And next, for a given decision situation, what is the best design for the test?

There is no procedure of which I am aware that is the best for all people under all situations and at all times. So what is best

depends on the person's values and judgement at the time when the decision is to be made. You have your own bias; so do I. Hence, there seems to be no way to avoid a Bayesian approach when formulating a psychometric theory for decisions based on test data. Such theory does require that you express the odds and ends of your values and prior bias in some numerical form, then resort to criteria such as maximum expected utility or minimum risk to solve the problem.

9. On the Interpretation of Test Scores: How Is Life Without a Prior?

For a Bayesian, life does not exist without a prior; but then, some of us may wish to express our values clearly and at the same time do not believe in any subjective probability. Well, you can certainly resort to optimizing criteria such as minimax. (For a devoted Bayesian, do not even mention this dirty word because it will force upon the decision problem the worst prior judgement.)

The minimax principle has been used successfully in many situations. Take the case of robust estimation and the Huber M-estimate: It is actually a minimax solution within the context of contaminated normal distributions.

However, life with minimax seems rather dull. You will take only the action which has the smallest maximum risk. But then think about life as a traveling experience: What happens if you only choose the road on which the mountains are not the highest and the rivers not the deepest? Some day, perhaps some of us will continue the job that Wald started many years ago and will devise a way to take care of personal values without prior judgement.

10. Epilogue

I hate to abruptly end the talk here, but the ways in which we may approach psychometrics are almost unlimited. Advances in computing technology, the extent to which test items are being shared across school district boundaries make possible a fresh look at some of the concepts in measurement which have been dear to us.

Some of the issues briefly referred to previously are not that simple nor easily solved. Perhaps they are not that important at all. But then, you never know.

We are in a country where freedom is our children's first word. But freedom carries along with it the uncomfortable notion of doubt. Having been quite sure about the first paper on mastery testing written at the University of Pittsburgh in the summer of 1973 and now at the conclusion of this final report to NIE, I just wonder whether educational testing or I myself have changed that much during those years. But I have the privilege of having a new generation of students every fall. Though winters sometimes have been harsh, springs always come with the early blossoms of dogwoods and daffodils. With the thoughts about these beautiful flowers and the fresh memory of all the students who have passed through my offices including VMC, EMH, LM, JCS, and JC, I am now at the very end of this report.